Appendix

A Proofs for Section 2

Proof of Lemma 2.1

Proof. We construct a "ghost" point:

$$x_1 = \mathcal{P}_{\mathcal{X}}\left(x - \frac{1}{\beta}\nabla_x \tilde{f}([x]_\beta, [y]_\beta)\right), \quad y_1 = \mathcal{P}_{\mathcal{Y}}\left(y + \frac{1}{\beta}\nabla_y \tilde{f}([x]_\beta, [y]_\beta)\right).$$

From (x, y) to (x_1, y_1) is just one step of extra-gradient with stepsize $\frac{1}{\beta_0}$. According to [34] or Section 4.5 of [4], we have

$$\begin{aligned}
&\nabla_x \tilde{f}([x]_\beta, [y]_\beta)^T([x]_\beta - \bar{x}) - \nabla_y \tilde{f}([x]_\beta, [y]_\beta)^T([y]_\beta - \bar{y}) \\
\leq &\frac{\beta}{2}[(\|x - \bar{x}\|^2 + \|y - \bar{y}\|^2) - (\|x_1 - \bar{x}\|^2 + \|y_1 - \bar{y}\|^2)], \quad \forall \bar{x} \in \mathcal{X}, \bar{y} \in \mathcal{Y}. \qquad (10)
\end{aligned}$$

1. Denote $x^*(y) = \arg\min_{x\in\mathcal{X}} \tilde{f}(x, y)$ and $y^*(x) = \arg\max_{y\in\mathcal{Y}} \tilde{f}(x, y)$. By convexity-concavity of $\tilde{f}$, we have

$$\begin{aligned}
\mathrm{gap}_{\tilde{f}}([z]_\beta) =&\tilde{f}([x]_\beta, [y]_\beta) - \min_{x\in\mathcal{X}} \tilde{f}(x, [y]_\beta) + \max_{y\in\mathcal{Y}} \tilde{f}([x]_\beta, y) - \tilde{f}([x]_\beta, [y]_\beta) \\
\leq&\nabla_x \tilde{f}([x]_\beta, [y]_\beta)^T([x]_\beta - x^*([y]_\beta)) - \nabla_y \tilde{f}([x]_\beta, y_{k+1/2})^T([y]_\beta - y^*([x]_\beta)) \\
\leq&\frac{\beta}{2}[(\|x - x^*([y]_\beta)\|^2 + \|y - y^*([x]_\beta)\|^2) - (\|x_1 - x^*([y]_\beta)\|^2 + \|y_1 - y^*([x]_\beta)\|^2)] \\
\leq&\beta[\|x - x^*\|^2 + \|x^* - x^*([y]_\beta)\|^2 + \|y - y^*\|^2 + \|y^* - y^*([x]_\beta)\|^2] \qquad (11) \\
\leq&\beta[\|x - x^*\|^2 + \|y - y^*\|^2] + \frac{\beta\tilde{\ell}^2}{\tilde{\mu}^2}[\|[x]_\beta - x^*\|^2 + \|[y]_\beta - y^*\|^2] \\
\leq&\left(\beta + \frac{2\beta\tilde{\ell}^2}{\tilde{\mu}^2}\right)[\|x - x^*\|^2 + \|y - y^*\|^2] + \frac{2\beta\tilde{\ell}^2}{\tilde{\mu}^2}[\|[x]_\beta - x\|^2 + \|[y]_\beta - y\|^2], \qquad (12)
\end{aligned}$$

where in the second inequality we apply (10) , in the third and last inequalities we use Young's inequality, and in the fourth inequality we use $\|x^* - x^*([y]_\beta)\| = \|x^*(y^*) - x^*([y]_\beta)\| \leq \frac{\tilde{\ell}}{\tilde{\mu}}\|[y]_\beta - y^*\|$ (and similarly for $\|y^* - y^*([x]_\beta)\|$, see Lemma B.2 in [25]). From Lemma 3.1 and Proposition 3.2 in [48], we have

$$\|[x]_\beta - x\|^2 + \|[y]_\beta - y\|^2 \leq \frac{1}{(1 - \tilde{\ell}/\beta)^2}[\|x - x_1\|^2 + \|y - y_1\|^2] \leq \frac{2}{(1 - \tilde{\ell}/\beta)^3}[\|x - x^*\|^2 + \|y - y^*\|^2]. \qquad (13)$$

Combining with (12), we have

$$\mathrm{gap}_{\tilde{f}}([z]_\beta) \leq \left(\beta + \frac{2\beta\tilde{\ell}^2}{\tilde{\mu}^2} + \frac{4\beta\tilde{\ell}^2}{\tilde{\mu}^2(1 - \tilde{\ell}/\beta)^3}\right)[\|x - x^*\|^2 + \|y - y^*\|^2]. \qquad (14)$$

Then again from (10), for any arbitrary $\bar{y} \in \mathcal{Y}$ we have

$$\begin{aligned}
&\nabla_x \tilde{f}([x]_\beta, [y]_\beta)^T([x]_\beta - x^*([y]_\beta)) - \nabla_y \tilde{f}([x]_\beta, [y]_\beta)^T([y]_\beta - y) \\
\leq&\frac{\beta}{2}[(\|x - x^*([y]_\beta)\|^2 + \|y - \bar{y}\|^2) - (\|x_1 - x^*([y]_\beta)\|^2 + \|y_1 - \bar{y}\|^2)] \\
\leq&\frac{\beta}{2}\|x - x^*([y]_\beta)\|^2 + \frac{\beta}{2}[\|y - \bar{y}\|^2 - \|y_1 - \bar{y}\|^2] \\
\leq&\frac{\beta}{2}\|x - x^*([y]_\beta)\|^2 + \frac{\beta}{2}\|y - y_1\|\|y - \bar{y} + y_1 - \bar{y}\| \\
\leq&\left(\beta + \frac{2\beta\tilde{\ell}^2}{\tilde{\mu}^2} + \frac{4\beta\tilde{\ell}^2}{\tilde{\mu}^2(1 - \tilde{\ell}/\beta)^3}\right)[\|x - x^*\|^2 + \|y - y^*\|^2] + \beta\mathcal{D}_{\mathcal{Y}}[\|y - y^*\| + \|y_1 - y^*\|] \\
\leq&\left(\beta + \frac{2\beta\tilde{\ell}^2}{\tilde{\mu}^2} + \frac{4\beta\tilde{\ell}^2}{\tilde{\mu}^2(1 - \tilde{\ell}/\beta)^3}\right)[\|x - x^*\|^2 + \|y - y^*\|^2] + 2\beta\mathcal{D}_{\mathcal{Y}}[\|x - x^*\| + \|y - y^*\|], \qquad (15)
\end{aligned}$$

where in the fourth inequality, we bound $\|x - x^*([y]_\beta)\|^2$ the same way as we did from (11) to (13), and in the last inequality we use $\|z - z^*\| \le \|z_1 - z^*\|$ (Proposition 3.2 in [48]). By noting that

$$\nabla_x \tilde{f}([x]_\beta, [y]_\beta)^T([x]_\beta - x^*([y]_\beta)) \ge 0,$$

we reach our conclusion.

2. Theorem 3.1 of [41] shows the relationship between $\|x-x^*\|+\|y-y^*\|$ and $\|x-[x]_\beta\|+\|y-[y]_\beta\|$ in the case $\beta = 1$. The proof can be extended to the following general case:

$$\|x - x^*\| + \|y - y^*\| \le \frac{\beta + \tilde{\ell}}{\tilde{\mu}}[\|x - [x]_\beta\| + \|y - [y]_\beta\|].$$

The last relationship we want to show is just equation (13). □

B Proofs for Section 3

Proof of Theorem 3.1

Proof. Because $\Phi_{x_t}(y) := \tilde{f}_t(x_t, y) = f(x_t, y) - \frac{\tau}{2}\|y - z_t\|^2$ is τ-strongly-concave, we have

$$\Phi_{x_t}(y_t) - \Phi_{x_t}(y) \ge \frac{1}{2}\tau\|y - y_t\|^2 + \nabla_y \tilde{f}(x_t, y_t)^T(y_t - y), \quad \forall y \in \mathcal{Y}.$$

With stopping criterion of the subproblem (3), we have

$$f(x_t, y_t) - f(x_t, y) \ge \frac{1}{2}\tau\|y - y_t\|^2 + \frac{\tau}{2}\|y_t - z_t\|^2 - \frac{\tau}{2}\|y - z_t\|^2 - \epsilon^{(t)}. \tag{16}$$

Choose $y = \alpha_t\tilde{y} + (1 - \alpha_t)y_{t-1}$ in (16), where $\tilde{y}$ is an arbitrary vector in $\mathcal{Y}$, then

$$f(x_t, \tilde{y})-f(x_t, y_t) \le (1-\alpha_t)[f(x_t, \tilde{y})-f(x_t, y_{t-1})]-\frac{\tau}{2}\alpha_t^2(\|v_t-\tilde{y}\|^2-\|v_{t-1}-\tilde{y}\|^2)-\frac{\tau}{2}\|y_t-z_t\|^2+\epsilon^{(t)}. \tag{17}$$

Note that

$$\begin{aligned} f(x_t, \tilde{y}) - f(x_t, y_{t-1}) =& f(x_{t-1}, \tilde{y}) - f(x_{t-1}, y_{t-1}) + f(x_{t-1}, y_{t-1}) - f(x_t, y_{t-1}) + f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y}) \\ \le& f(x_{t-1}, \tilde{y}) - f(x_{t-1}, y_{t-1}) + f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y}) + \epsilon^{(t-1)}, \end{aligned} \tag{18}$$

where the inequality follows because $f(x_t, y_t) - \min_{x\in\mathcal{X}} f(x, y_t) \le \epsilon^{(t)}$. Plugging this back to (17) and rearranging,

$$\begin{aligned} \frac{1}{\alpha_t^2}[f(x_t, \tilde{y}) - f(x_t, y_t)] + \frac{\tau}{2}\|v_t - \tilde{y}\|^2 \le& \frac{1 - \alpha_t}{\alpha_t^2}[f(x_{t-1}, \tilde{y}) - f(x_{t-1}, y_{t-1})] + \frac{\tau}{2}\|v_{t-1} - \tilde{y}\|^2 + \\ & \frac{1 - \alpha_t}{\alpha_t^2}[f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y})] + \frac{1 - \alpha_t}{\alpha_t^2}\epsilon^{(t-1)} + \frac{1}{\alpha_t^2}\epsilon^{(t)}. \end{aligned} \tag{19}$$

Using the update rule for sequence $\{\alpha_t\}_t$, for $t > 1$ we have

$$\begin{aligned} \frac{1}{\alpha_t^2}[f(x_t, \tilde{y}) - f(x_t, y_t)] + \frac{\tau}{2}\|v_t - \tilde{y}\|^2 \le& \frac{1}{\alpha_{t-1}^2}[f(x_{t-1}, \tilde{y}) - f(x_{t-1}, y_{t-1})] + \frac{\tau}{2}\|v_{t-1} - \tilde{y}\|^2 + \\ & \frac{1}{\alpha_{t-1}^2}[f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y})] + \frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \frac{1}{\alpha_t^2}\epsilon^{(t)}. \end{aligned} \tag{20}$$

Iterating this inequality results in

$$\begin{aligned} \frac{1}{\alpha_t^2}[f(x_t, \tilde{y}) - f(x_t, y_t)] + \frac{\tau}{2}\|v_t - \tilde{y}\|^2 \le& \frac{1}{\alpha_1^2}[f(x_1, \tilde{y}) - f(x_1, y_1)] + \frac{\tau}{2}\|v_1 - \tilde{y}\|^2 + \\ & \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}[f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y})] + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \sum_{t=2}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)} \\ =& f(x_1, \tilde{y}) - f(x_1, y_1) + \frac{\tau}{2}\|v_1 - \tilde{y}\|^2 + \\ & \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}[f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y})] + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \sum_{t=2}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)}, \end{aligned} \tag{21}$$

where we use $\alpha_1 = 1$. Applying (19) with $t = 1$ (note $\alpha_1 = 1$), we have

$$f(x_1, \tilde{y}) - f(x_1, y_1) + \frac{\tau}{2}\|v_1 - \tilde{y}\|^2 \le \frac{\tau}{2}\|y_0 - \tilde{y}\|^2 + \epsilon^{(1)}. \tag{22}$$

Combining with (21),

$$\begin{aligned}
&\frac{1}{\alpha_T^2}[f(x_T, \tilde{y}) - f(x_T, y_T)] + \frac{\tau}{2}\|v_T - \tilde{y}\|^2 \\
\le& \frac{\tau}{2}\|y_0 - \tilde{y}\|^2 + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}[f(x_t, \tilde{y}) - f(x_{t-1}, \tilde{y}] + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)} \\
\le& \frac{\tau}{2}\|y_0 - \tilde{y}\|^2 + \frac{1}{\alpha_{T-1}^2}f(x_T, \tilde{y}) - \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}}f(x_{t-1}, \tilde{y}) + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)},
\end{aligned}$$

where in the last inequality we use $\frac{1}{\alpha_t^2} - \frac{1}{\alpha_{t-1}^2} = \frac{1}{\alpha_t}$. Rearranging,

$$\begin{aligned}
&\frac{\tau}{2}\|y_0 - \tilde{y}\|^2 + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}^2}\epsilon^{(t-1)} + \sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)} \\
\ge& \frac{1}{\alpha_T^2}[f(x_T, \tilde{y}) - f(x_T, y_T)] + \frac{\tau}{2}\|v_T - \tilde{y}\|^2 - \frac{1}{\alpha_{T-1}^2}f(x_T, \tilde{y}) + \sum_{t=2}^{T}\frac{1}{\alpha_{t-1}}f(x_{t-1}, \tilde{y}) \\
\ge& \sum_{t=1}^{T}\frac{1}{\alpha_t}f(x_t, \tilde{y}) - \frac{1}{\alpha_T^2}f(x_T, y_T) \\
\ge& \sum_{m=1}^{T}\frac{1}{\alpha_m}f\left(\sum_{t=1}^{T}\frac{1/\alpha_t}{\sum_{k=1}^{T}1/\alpha_k}x_t, \tilde{y}\right) - \frac{1}{\alpha_T^2}f(x_T, y_T) \\
\ge& \sum_{m=1}^{T}\frac{1}{\alpha_m}f\left(\sum_{t=1}^{T}\frac{1/\alpha_t}{\sum_{k=1}^{T}1/\alpha_k}x_t, \tilde{y}\right) - \frac{1}{\alpha_T^2}f(\tilde{x}, y_T) - \frac{1}{\alpha_T^2}\epsilon^{(T)}, \quad \forall \tilde{x} \in \mathcal{X},
\end{aligned}$$

where in the third inequality we use the convexity of $f(\cdot, \tilde{y})$, and in the last inequality we use $f(x_t, y_t) - \min_{x\in\mathcal{X}} f(x, y_t) \le \epsilon^{(t)}$. Note that

$$\sum_{m=1}^{t}\frac{1}{\alpha_m} = \frac{1}{\alpha_1} + \left(\frac{1}{\alpha_2^2} - \frac{1}{\alpha_1^2}\right) + \left(\frac{1}{a_3^2} - \frac{1}{\alpha_2^2}\right) + ... + \left(\frac{1}{\alpha_t^2} - \frac{1}{\alpha_{t-1}^2}\right) = \frac{1}{\alpha_t^2}. \tag{23}$$

Therefore

$$f(\bar{x}_T, \tilde{y}) - f(\tilde{x}, y_T) \le a_T^2\left[\frac{\tau}{2}\|y_0 - \tilde{y}\|^2 + 2\sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)}\right], \quad \forall \tilde{x} \in \mathcal{X}, \tilde{y} \in \mathcal{Y}, \tag{24}$$

which directly implies

$$\text{gap}_f(\bar{x}_T, y_T) \le \alpha_T^2\left[\frac{\tau}{2}\mathcal{D}_{\mathcal{Y}}^2 + 2\sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)}\right]. \tag{25}$$

By choosing $\epsilon^{(t)} = \frac{3\tau\mathcal{D}_{\mathcal{Y}}\alpha_t^2}{2\pi t^2}$,

$$\sum_{t=1}^{T}\frac{1}{\alpha_t^2}\epsilon^{(t)} = \frac{3\tau\mathcal{D}_{\mathcal{Y}}}{2\pi}\sum_{t=1}^{T}\frac{1}{t^2} \le \frac{\tau\mathcal{D}_{\mathcal{Y}}}{4}, \tag{26}$$

therefore,

$$\text{gap}_f(\bar{x}_T, y_T) \le \alpha_T^2\tau\mathcal{D}_{\mathcal{Y}}^2. \tag{27}$$

□

Proof of Proposition 3.1

Proof. First, we show that the initial point (x_{t-1}, z_t) will not be infinitely far from the saddle point (x_t^*, y_t^*) of the subproblem $(\star)$ at t-th iteration of outer loop. Since $\mathcal{Y}$ is bounded, we have $\|z_t - y_t^*\| \le \mathcal{D}_{\mathcal{Y}}$. Denote $x^*(y) = \text{argmin}_x f(x, y)$. Since $f(\cdot, y)$ is μ-strongly convex, we have

$$\|x^*(y_{t-1}) - x^*(y_t)\| \le \frac{l}{\mu}\|y_t - y_{t-1}\| \le \frac{l}{\mu}\mathcal{D}_{\mathcal{Y}}, \tag{28}$$

where we use Lemma B.2 in [25]. Further with the strong convexity of $f(\cdot, y_{t-1})$, we have

$$\|x_{t-1} - x_t^*\|^2 \leq 2\|x_{t-1} - x^*(y_{t-1}^*)\|^2 + 2\|x^*(y_{t-1}^*) - x^*(y_t^*)\|^2 \leq \frac{4\epsilon^{(t-1)}}{\mu_x} + 2\left(\frac{l}{\mu_x}\right)^2 \mathcal{D}_{\mathcal{Y}}.$$

Therefore, the distance from the initial point to the saddle point of the subproblem is bounded. From now, we use subscript to index the iteration of the inner-loop and (x_0, y_0) denotes the initial point we specified above. We separate the discussion into deterministic and stochastic settings.

Deterministic setting. We apply a deterministic algorithm $\mathcal{M}$ to solve the subproblem and $\mathcal{M}$ has a linear rate described by (4). By Lemma 2.1, after K iterations of algorithm $\mathcal{M}$,

$$\begin{aligned}\|x_K - [x_K]_\beta\|^2 + \|y_K - [y_K]_\beta\|^2 \leq & \frac{2}{(1-\tilde{\ell}/\beta)^3}[\|x_K - x^*\|^2 + \|y_K - y^*\|^2] \\ & \leq \frac{2}{(1-\tilde{\ell}/\beta)^3}\left(1 - \frac{1}{\Delta_{\mathcal{M},\tau}}\right)^K [\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2].\end{aligned}$$

Choosing

$$K = \Delta_{\mathcal{M},\tau} \log \frac{(1-\tilde{\ell}/\beta)^3(\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2)}{2\epsilon},$$

we have $\|x_K - [x_K]_\beta\|^2 + \|y_K - [y_K]_\beta\|^2 \leq \epsilon$. To satisfy condition (6), it suffices to set

$$\epsilon = \min\left\{\frac{\tilde{\mu}^2\epsilon^{(t)}}{2A(\beta+\tilde{\ell})^2}, \left(\frac{\tilde{\mu}\epsilon^{(t)}}{4\beta\mathcal{D}_{\mathcal{Y}}(\beta+\tilde{\ell})}\right)^2\right\},$$

and we reach our conclusion.

Stochastic setting. We apply a stochastic algorithm $\mathcal{M}$ to solve the subproblem and $\mathcal{M}$ has a linear rate described by (5). With the same reasoning as in deterministic setting and applying Appendix B.4 of [23], we have

$$K(\epsilon) \leq \Delta_{\mathcal{M},\tau} \log \frac{(1-\tilde{\ell}/\beta)^3(\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2)}{2\Delta_{\mathcal{M},\tau}\epsilon} + 1,$$

and the conclusion follows directly.

□

Proof of Corollary 3.2

Proof. Because $2/(t+2)^2 \leq \alpha_t^2 \leq 4/(t+1)^2$, by Theorem 3.1, Algorithm 1 finds ϵ-saddle point after $T = \mathcal{O}\left(\sqrt{\mu/\epsilon} \cdot \mathcal{D}_{\mathcal{Y}}\right)$ outer-loop iterations. Note that the accuracy we want for solving subproblem $(\star)$ is

$$\epsilon^{(t)} = \frac{3\tau\mathcal{D}_{\mathcal{Y}}\alpha_t^2}{2\pi t^2} \geq \frac{6\tau\mathcal{D}_{\mathcal{Y}}}{\pi t^2(t+2)^2} \geq \frac{6\tau\mathcal{D}_{\mathcal{Y}}}{\pi T^2(T+2)^2} = \Omega(\epsilon^2\mu^{-1}\mathcal{D}_{\mathcal{Y}}^{-3}), \quad \forall t \in [T]. \tag{29}$$

By Proposition 3.1, it takes at most

$$K = \mathcal{O}\left(\Delta_{\mathcal{M},\tau} \log\left(\frac{\ell\mathcal{D}_{\mathcal{Y}}}{\min\{1, \mu, \tau\}\epsilon}\right)\right)$$

gradient oracle calls for $\mathcal{M}$ to solve the subproblem. The total complexity is then $K \cdot T$.

□

C Proofs for Section 4

Proof of Theorem 4.1

Proof. First we define ψ as the extended-value function of g: $\psi(x) = g(x)$ if $x \in \mathcal{X}$ and $\psi(x) = \infty$ if $x \notin \mathcal{X}$. Note that $g(x) = \max_{y\in\mathcal{Y}} f(x, y)$ is ℓ-weakly convex [Lemma 3, [47]]. It directly follows from the definition of ψ that ψ is also ℓ-weakly convex. Define the proximal point of x by

$$\text{prox}_{\tau\psi}(x) = \text{argmin}_z \left\{\psi(z) + \frac{1}{2\tau}\|z - x\|^2\right\} = \text{argmin}_{z\in\mathcal{X}}\, g_{1/\tau}(z; x).$$

By [Lemma 4.3 in [11]], as $\tau_x > \ell$,

$$\begin{aligned}\|\nabla\psi_{1/\tau_x}(x_t)\|^2 = \tau_x^2\|x_t - \mathrm{prox}_{\psi/\tau_x}(x_t)\|^2 \leq & \frac{2\tau_x^2}{\tau_x - \ell}[g_{\tau_x}(x_t; x_t) - g_{\tau_x}(\mathrm{prox}_{\psi/\tau_x}(x); x_t)] \\ \leq & \frac{2\tau_x^2}{\tau_x - \ell}[g_{\tau_x}(x_t; x_t) - g_{\tau_x}(x_{t+1}; x_t) + \bar{\epsilon}] \\ = & \frac{2\tau_x^2}{\tau_x - \ell}\left\{g(x_t) - \left[g(x_{t+1}) + \frac{\tau_x}{2}\|x_{t+1} - x_t\|^2\right] + \bar{\epsilon}\right\} \\ \leq & \frac{2\tau_x^2}{\tau_x - \ell}[g(x_t) - g(x_{t+1}) + \bar{\epsilon}], \end{aligned} \tag{30}$$

where in the first inequality we use $(\tau_x - \ell)$-strong convexity of $g_{\tau_x}(\cdot; x_t)$, and the second inequality follows from $g_{\tau_x}(x_{t+1}; x_t) \leq \min_{x\in\mathcal{X}} g_{\tau_x}(x; x_t) + \bar{\epsilon}$. Summing from 0 to $T-1$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\psi_{\tau_x}(x_t)\|^2 \leq \frac{2\tau_x^2}{\tau_x - \ell}\left[\frac{g(x_0) - g(x_T)}{T} + \bar{\epsilon}\right] \leq \frac{2\tau_x^2}{\tau_x - \ell}\left[\frac{g(x_0) - g^*}{T} + \bar{\epsilon}\right]. \tag{31}$$

□

Proof of Corollary 4.2

Proof. According to Theorem 4.1, with $\tau_x = 2\ell$, it takes at most $T = \frac{4\tau_x^2(g(x_0)-g^*)}{(\tau_x-\ell)\epsilon^2} = \frac{16\ell(g(x_0)-g^*)}{\epsilon^2}$ outer-loops to find ϵ-stationary point. The auxiliary problem $\min_{x\in\mathcal{X}} g_{\tau_x}(x; x_t)$ is then ℓ-SC-C and (3ℓ)-smooth. By Corollary 3.2 and discussion in Section 3.2, Algorithm 1 combined with EG/OGDA/GDA can solve such auxiliary problem with complexity $\tilde{\mathcal{O}}(\sqrt{\ell/\bar{\epsilon}}) = \tilde{\mathcal{O}}(\ell/\epsilon)$ as $\bar{\epsilon} = \frac{\epsilon^2}{8\ell}$ specified in Theorem 4.1. So the total complexity is $\tilde{\mathcal{O}}(\ell^2/\epsilon^3)$. □

Proof of Corollary 4.3

Proof. As we assume each f_i has ℓ-Lipschitz gradient, $f(x, y) = \frac{1}{n}\sum_{i=1}^n f_i(x, y)$ has $\bar{\ell}$-Lipschitz gradient. According to Theorem 4.1, with $\tau_x = 2\bar{\ell}$, it takes at most $T = \frac{4\tau_x^2(g(x_0)-g^*)}{(\tau_x-\ell)\epsilon^2} = \frac{16\bar{\ell}(g(x_0)-g^*)}{\epsilon^2}$ outer-loops to find ϵ-stationary point. The resulting auxiliary problem is $\bar{\ell}$-SC-C and $(3\bar{\ell})$-smooth. By Corollary 3.2, Algorithm 1 combined with EG/OGDA can solve such auxiliary problem with complexity

$$\tilde{\mathcal{O}}\left(\left(n + \left(\frac{3\bar{\ell} + \tau_y}{\min\{\bar{\ell}, \tau_y\}}\right)^2\right)\sqrt{\frac{\tau_y}{\bar{\epsilon}}}\right).$$

Choosing $\tau_y = \bar{\ell}/\sqrt{n}$ and $\bar{\epsilon} = \frac{\epsilon^2}{8\bar{\ell}}$, Algorithm 1 has complexity of $\tilde{\mathcal{O}}\left(n^{\frac{3}{4}}\bar{\ell}/\epsilon\right)$ to solve the auxiliary problem. The total complexity is therefore $\tilde{\mathcal{O}}\left(n^{\frac{3}{4}}\bar{\ell}^2\epsilon^{-3}\right)$.

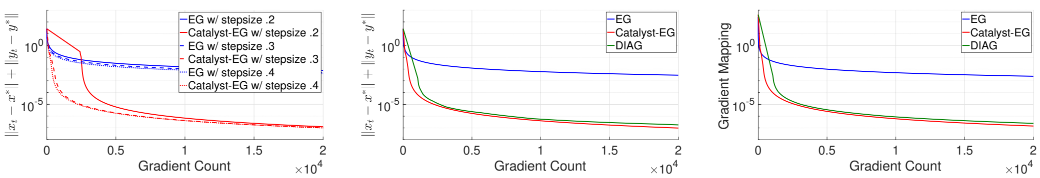
When we further assume f has ℓ_i-cocoercive gradient, Algorithm 1 combined with SVRE can solve such auxiliary problem with complexity

$$\tilde{\mathcal{O}}\left(\left(n + \frac{3\bar{\ell} + \tau_y}{\min\{\bar{\ell}, \tau_y\}}\right)\sqrt{\frac{\tau_y}{\bar{\epsilon}}}\right).$$

Choosing $\tau_y = \bar{\ell}/n$ and $\bar{\epsilon} = \frac{\epsilon^2}{8\bar{\ell}}$, Algorithm 1 has complexity of $\tilde{\mathcal{O}}\left(n^{\frac{1}{2}}\bar{\ell}/\epsilon\right)$ to solve the auxiliary problem. The total complexity is therefore $\tilde{\mathcal{O}}\left(n^{\frac{1}{2}}\bar{\ell}^2\epsilon^{-3}\right)$. □

D Additional Experiments

In this section, we provide additional experiments on SC-C minimax problems to illustrate the performance of Catalyst framework. Here we focus on the comparison between the performance of EG, Catalyst-EG and DIAG [47]. We implement these algorithms in the same way as in Section 5.



(a) Distance to limit point (b) Comparion of EG and Catalyst-EG (c) Norm of gradient mapping

Figure 4: SC-C experiment on distributionally robust logistic regression

D.1 Distributionally robust logistic regression

We consider the distributionally robust logistic regression problem [32]. This results in a minimax problem:

$$\min_{\theta} \max_{p \in \Delta_n} \sum_{i=1}^{n} -p_i \left[y_i \log\left(\hat{y}\left(X_i\right)\right) + \left(1-y_i\right) \log\left(1-\hat{y}\left(X_i\right)\right)\right] \text{ such that } \|p - \mathbf{1}/n\| \leq \rho, \quad (32)$$

where θ parametrizes the classifier $\hat{y}(\cdot)$, and (y, X) is classification data. When $\hat{y}(x) = \frac{e^{\theta^\top x}}{1+e^{\theta^\top x}}$, it can be formulated as the following SC-C minimax problem:

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \log(1 + \exp(-y_i \theta^\top X_i)) + \frac{\lambda}{2}\|\theta\|^2 \text{ such that } \|p - \mathbf{1}/n\| \leq \rho, \quad (33)$$

where λ is a regularization parameter.

We conduct experiments on the Wisconsin breast cancer dataset [13], which has 30 attributes and 569 samples. We separate 80% of the data as our training set. We compare the performance of EG, Catalyst-EG and DIAG. We compare EG and Catalyst-EG under same stepsizes in Figure 4(a). We also report two different error measures under the best-tuned stepsizes in Figure 4(b) and 4(c). We observe that Catalyst-EG performs consistently well. As algorithms designed for SC-C setting, both DIAG and Catalyst-EG converge faster than EG.