

A Technical Appendices and Supplementary Material

A.1 Analysis of the maximization variables

Consider the following parametric maximization problem

$$\tilde{h}^*(x) \triangleq \max_{u \in U} h(x, u), \quad (14)$$

for a given $x \in \mathbb{R}^n$ where $h(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a continuously differentiable function, and $U \subseteq \mathbb{R}^m$ is a closed, convex set. We are interested in the conditions under which the value function $\tilde{h}^*(x)$ is Lipschitz differentiable and an approximate solution of the above problem can be found using (accelerated) gradient ascent method. Indeed, using the classical result in the optimization literature [32, 56, 55, 57], it can be deduced that both these properties are satisfied when $h(x, \cdot)$ is strongly concave or satisfies PL inequality when $U = \mathbb{R}^m$. In the following, we will restate these results in a unified statement and later specify them for our proposed algorithm. In particular, we first state the linear convergence result under these conditions, and then state the differentiability of the value function.

Proposition A.1 ([32, 56, 55]). *Consider problem (14), and assume that $h(\cdot, \cdot)$ is a continuously differentiable function, such that for any fixed x , $h(x, \cdot)$ is either strongly concave (or satisfies PL inequality (see Def. 3 with $\theta = \frac{1}{2}$) with $U = \mathbb{R}^m$). Let $\{u_k\}_{k=0}^{T-1} \subset \mathbb{R}^m$ be a sequence generated by the (accelerated) gradient ascent method. Then, for any $x \in \mathbb{R}^n$, there exists $\delta \in (0, 1)$ and $\Delta_1, \Delta_2 > 0$ such that we have the following results for $T \geq 1$,*

$$\begin{aligned} \|u_T - u^*(x)\|^2 &\leq \Delta_1 \delta^T, \\ h(x, u_T) &\leq \tilde{h}^*(x) \leq h(x, u_T) + \Delta_2 \delta^T, \end{aligned}$$

where $u^*(x) \triangleq \mathcal{P}_{U^*(x)}(u_T)$, and $U^*(x) \triangleq \operatorname{argmax}_{u \in U} h(x, u)$.

Proposition A.2 ([57] Lemma A.5). *Consider problem (14), and assume that $h(\cdot, \cdot)$ is a continuously differentiable function, such that for any fixed x , $h(x, \cdot)$ is η_h -strongly concave (or c_h -PL (see Def. 2.1 with $\theta = \frac{1}{2}$) with $U = \mathbb{R}^m$), it follows that*

$$\nabla \tilde{h}^*(x) = \nabla_x h(x, u^*) \quad \text{for any } u^* \in U^*(x).$$

where $U^*(x) \triangleq \operatorname{argmax}_{u \in U} h(x, u)$. Moreover, $\tilde{h}^*(x)$ has $(L_{uu}^h + (L_{xu}^h)^2/\iota)$ -Lipschitz gradient where $\iota = \eta_h$ (or $\iota = c_h^2$).

Now, we apply the above propositions to the objective and constraint functions in problem (1) to derive the error of estimating the maximization components according to the updates of Algorithm 1, which will be used in the analysis. Recall that $f(x) = \max_{y \in Y} \phi(x, y)$ and $g(x) = \max_{w \in W} \psi(x, w)$. Based on Proposition A.2 and Assumptions 2.1 and 2.2, we have the following properties:

1. f is continuously differentiable and has a Lipschitz gradient with constant $L_f \triangleq L_{yy}^\phi + (L_{xy}^\phi)^2/\iota_f$,
2. g is continuously differentiable and has a Lipschitz gradient with constant $L_g \triangleq L_{ww}^\psi + (L_{xw}^\psi)^2/\iota_g$,

where $\iota_f = \eta_\psi$ when $\psi(x, \cdot)$ is η_ψ -strongly concave or $\iota_f = c_\psi^2$ when $\psi(x, \cdot)$ is c_ψ -PL (ι_g is defined similarly).

Moreover, based on Proposition A.1, there exist uniform constants $\Delta_1^y, \Delta_1^w, \Delta_2^w \in (0, +\infty)$ and $\delta_y, \delta_w \in (0, 1)$, such that for any $k \geq 0$,

$$\|y_k - y^*(x_k)\|^2 \leq \Delta_1^y \delta_y^{N_k}, \quad (15)$$

$$\|w_k - w^*(x_k)\|^2 \leq \Delta_1^w \delta_w^{M_k}, \quad (16)$$

$$\psi(x_k, w_k) \leq g(x_k) \leq \psi(x_k, w_k) + \Delta_2^w \delta_w^{M_k}, \quad (17)$$

where N_k, M_k denote the number of (accelerated) gradient ascent steps to maximize the functions $\phi(x, \cdot)$ and $\psi(x, \cdot)$, respectively.

A.2 Required Lemmas

This section states two important lemmas regarding the proposed method and the implicit constraint function. First, we show some bounds on the modified dual multiplier λ_k corresponding to the subproblem (7) based on the update of Algorithm 1. Next, we establish local Lipschitz continuity of the infeasibility residual function $p(x) \triangleq [g(x)]_+^2$. Later, we show that by carefully selecting the stepsize, this local constant can be upper bounded by a global one.

Lemma A.3. *Suppose Assumptions 2.1 and 2.2 hold, and let $\{x_k, \lambda_k\}_{k \geq 0}$ be the sequence generated by Algorithm 1 such that $\{\alpha_k\}_{k \geq 0} \subset \mathbb{R}_+$ is non-increasing sequence. Then, for any $k \geq 0$ we have that $\lambda_k \rho(x_k, w_k) \leq \|\nabla_x \phi(x_k, y_k)\| + \alpha_k$. Furthermore, if Assumption 2.3 hold, then for any $k \geq 0$, $\lambda_k [g(x_k)]_+ \leq C[g(x_k)]_+^{2-2\theta} + C\Delta_2^w \delta_w^{M_k} [\psi(x_k, w_k)]_+^{1-2\theta}$ where $C \triangleq \mu(C_\phi + \alpha_0)$.*

Proof. Recall that $\rho(x_k, w_k) = \|\nabla_x \psi(x_k, w_k)\|$ and $\zeta(x_k, w_k) = [\psi(x_k, w_k)]_+ \|\nabla_x \psi(x_k, w_k)\|$. Note that if $\lambda_k = 0$, the bound holds trivially. Now suppose, $\zeta(x_k, w_k) > 0$, then using the update of λ_k we have that

$$\lambda_k \rho(x_k, w_k) = \frac{1}{\|\nabla_x \psi(x_k, w_k)\|} [-\nabla_x \psi(x_k, w_k)^\top \nabla_x \phi(x_k, y_k) + \alpha_k \rho(x_k, w_k)]_+.$$

Taking the absolute value from both sides, using the fact that $|\max\{a, b\}| \leq |a| + |b|$ for any $a, b \in \mathbb{R}$, followed by the triangle and Cauchy-Schwarz inequalities, we conclude that $\lambda_k \|\nabla_x \psi(x_k, w_k)\| \leq \|\nabla_x \phi(x_k, y_k)\| + \alpha_k$.

Similarly, from the definition of λ_k and Assumption 2.3 we conclude that

$$\begin{aligned} \lambda_k [g(x_k)]_+ &\leq \frac{[g(x_k)]_+}{\|\nabla_x \psi(x_k, w_k)\|} (\|\nabla_x \phi(x_k, y_k)\| + \alpha_k) \\ &\leq \mu [g(x_k)]_+ [\psi(x_k, w_k)]_+^{1-2\theta} (\|\nabla_x \phi(x_k, y_k)\| + \alpha_k) \\ &\leq \mu ([\psi(x_k, w_k)]_+^{2-2\theta} + \Delta_2^w \delta_w^{M_k} [\psi(x_k, w_k)]_+^{1-2\theta}) (\|\nabla_x \phi(x_k, y_k)\| + \alpha_k) \\ &\leq \mu ([g(x_k)]_+^{2-2\theta} + \Delta_2^w \delta_w^{M_k} [\psi(x_k, w_k)]_+^{1-2\theta}) (\|\nabla_x \phi(x_k, y_k)\| + \alpha_k) \end{aligned}$$

where in the penultimate inequality we used the second inequality in (17) and that α_k is a non-increasing sequence. The last inequality above follows from the first inequality in (17). Finally, the result follows from the boundedness of $\nabla_x \phi(\cdot, \cdot)$ – see Assumption 2.1. \square

Lemma A.4. *Suppose Assumption 2.2 holds. Let $g(x) \triangleq \max_{w \in W} \psi(x, w)$ and the infeasibility residual by $p(x) = [g(x)]_+^2$. Then $p(\cdot)$ is a continuously differentiable function and $\nabla p(x)$ is locally Lipschitz continuous with constant $L_p(x) \triangleq 2C_\psi^2 + L_g^2 + [g(x)]_+^2$.*

Proof. Differentiability of $p(\cdot)$ follows from differentiability of g as established in Property 2 and its gradient can be calculated by the chain rule as $\nabla p(x) = 2\nabla g(x)[g(x)]_+$. Therefore, we have that

$$\begin{aligned} \|\nabla p(x) - \nabla p(y)\| &= \|2[g(x)]_+ \nabla g(x) - 2[g(y)]_+ \nabla g(y)\| \\ &= \|2[g(x)]_+ (\nabla g(x) - \nabla g(y)) + 2\nabla g(y)([g(x)]_+ - [g(y)]_+)\| \\ &\leq 2\|\nabla g(x) - \nabla g(y)\| \| [g(x)]_+ \| + 2\|\nabla g(y)\| \| [g(x)]_+ - [g(y)]_+ \|, \end{aligned}$$

where in the second equality we added and subtracted $2[g(x)]_+ \nabla g(y)$. Note that based on Assumption 2.2(ii), we have that $\nabla g(x) = \nabla_x \psi(x, w^*(x))$ is bounded by C_ψ , hence, g is C_ψ -Lipschitz continuous. Therefore,

$$\|\nabla p(x) - \nabla p(y)\| \leq (2L_g [g(x)]_+ + 2C_\psi^2) \|x - y\|.$$

From Young's inequality, we can bound $2L_g [g(x)]_+ \leq L_g^2 + [g(x)]_+^2$. Therefore, the following holds

$$\|\nabla p(x) - \nabla p(y)\| \leq (2C_\psi^2 + L_g^2 + [g(x)]_+^2) \|x - y\|.$$

\square

A.3 Proof of one-step analysis

In this section, we prove the one-step analysis for the objective and constraints.

Lemma A.5. Suppose Assumptions 2.1, 2.2, and 2.3 hold. Let $\{x_k, \lambda_k\}_{k \geq 0}$ be the sequence generated by Algorithm 1 such that $\{\alpha_k\}_k$ is a non-increasing sequence and $\gamma_k \leq (L_f + L_{xy}^\phi)^{-1}$. Then, for any $k \geq 0$

$$(I) \quad \frac{\gamma_k}{2} \|d_k\|^2 \leq f(x_k) - f(x_{k+1}) + \gamma_k \alpha_k (C_\phi + \alpha_k) + \frac{L_{xy}^\phi}{2} \Delta_1^y \delta_y^{N_k}, \quad (18)$$

$$(II) \quad \gamma_k \alpha_k \zeta(x_k, w_k) \leq \frac{p(x_k)}{2} - \frac{p(x_{k+1})}{2} + \gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) \\ + \frac{L_{xw}^\psi}{2} p(x_k) \Delta_1^w \delta_w^{M_k} + \frac{\gamma_k^2 (L_p(x_k) + 2L_{xw}^\psi)}{4} \|d_k\|^2. \quad (19)$$

Proof. **Part (I):** Using Lipschitz continuity of gradient of f as established in Property 1 and update of x_{k+1} , we have that

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L_f}{2} \|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \gamma_k \nabla f(x_k)^\top d_k + \frac{\gamma_k^2 L_f}{2} \|d_k\|^2 \\ &= f(x_k) + \gamma_k (\nabla_x \phi(x_k, y_k) + d_k)^\top d_k + \left(\frac{\gamma_k^2 L_f}{2} - \gamma_k \right) \|d_k\|^2 \\ &\quad + \gamma_k \|\nabla f(x_k) - \nabla_x \phi(x_k, y_k)\| \|d_k\| \\ &\leq f(x_k) - \gamma_k \lambda_k \nabla_x \psi(x_k, w_k)^\top d_k + \left(\frac{\gamma_k^2 L_f}{2} - \gamma_k \right) \|d_k\|^2 + \gamma_k L_{xy}^\phi \|y_k - y^*(x_k)\| \|d_k\|, \end{aligned}$$

where in the last inequality we used $d_k = -\nabla_x \phi(x_k, y_k) - \lambda_k \nabla_x \psi(x_k, w_k)$, $\nabla f(x_k) = \nabla_x \phi(x_k, y^*(x_k))$, and Lipschitz continuity of the gradient of function ϕ . Moreover, from complementarity slackness condition we know that $\lambda_k (\nabla_x \psi(x_k, w_k)^\top d_k + \alpha_k \rho(x_k, w_k)) = 0$, hence we obtain

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \left(\frac{\gamma_k^2 L_f}{2} - \gamma_k \right) \|d_k\|^2 + \gamma_k \alpha_k \lambda_k \rho(x_k, w_k) + \gamma_k L_{xy}^\phi \|y_k - y^*(x_k)\| \|d_k\| \\ &\leq \left(\frac{\gamma_k^2 L_f}{2} - \gamma_k \right) \|d_k\|^2 + \gamma_k \alpha_k (C_\phi + \alpha_k) + \gamma_k L_{xy}^\phi \|y_k - y^*(x_k)\| \|d_k\| \\ &\leq \left(\frac{\gamma_k^2 L_f}{2} - \gamma_k \right) \|d_k\|^2 + \gamma_k \alpha_k (C_\phi + \alpha_k) + \frac{L_{xy}^\phi}{2} \|y_k - y^*(x_k)\|^2 + \frac{\gamma_k^2 L_{xy}^\phi}{2} \|d_k\|^2. \quad (20) \end{aligned}$$

where the penultimate inequality follows from the application of Lemma A.3 and $\|\nabla_x \phi(x, y)\| \leq C_\phi$, moreover, the last inequality is due to Young's inequality (where $p = q = 2$). Now, rearranging the terms and selecting $\gamma_k \leq (L_f + L_{xy}^\phi)^{-1}$ lead to the result of part (I).

Part (II): Recall that $\zeta(x_k, w_k) = [\psi(x_k, w_k)]_+ \|\nabla_x \psi(x_k, w_k)\|$ and $\rho(x_k, w_k) = \|\nabla_x \psi(x_k, w_k)\|$. Based on Lemma A.4 and the update rule of $x_{k+1} = x_k + \gamma_k d_k$, we have that

$$\begin{aligned} p(x_{k+1}) - p(x_k) &\leq \langle \nabla p(x_k), x_{k+1} - x_k \rangle + \frac{L_p(x_k)}{2} \|x_{k+1} - x_k\|^2 \\ &= 2\gamma_k [g(x_k)]_+ \nabla g(x_k)^\top d_k + \frac{\gamma_k^2 L_p(x_k)}{2} \|d_k\|^2 \\ &= 2\gamma_k [g(x_k)]_+ \nabla_x \psi(x_k, w_k)^\top d_k + 2\gamma_k [g(x_k)]_+ (\nabla g(x_k) - \nabla_x \psi(x_k, w_k))^\top d_k + \frac{\gamma_k^2 L_p(x_k)}{2} \|d_k\|^2 \\ &\leq 2\gamma_k \underbrace{[g(x_k)]_+ \nabla_x \psi(x_k, w_k)^\top d_k}_{\text{term (a)}} + 2\gamma_k L_{xy}^\psi [g(x_k)]_+ \|w_k - w^*(x_k)\| \|d_k\| + \frac{\gamma_k^2 L_p(x_k)}{2} \|d_k\|^2. \quad (21) \end{aligned}$$

Considering term (a), from (17) one can observe that

$$\begin{aligned}
[g(x_k)]_+ \nabla_x \psi(x_k, w_k)^\top d_k &\leq [\psi(x_k, w_k)]_+ \nabla_x \psi(x_k, w_k)^\top d_k + \Delta_1^w \delta_w^{M_k} \|\nabla_x \psi(x_k, w_k)\| \|d_k\| \\
&\leq -\alpha_k [\psi(x_k, w_k)]_+ \rho(x_k, w_k) + \Delta_1^w \delta_w^{M_k} \|\nabla_x \psi(x_k, w_k)\| \|d_k\| \\
&\leq -\alpha_k \zeta(x_k, w_k) + \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0),
\end{aligned} \tag{22}$$

where in the second inequality we use the fact that d_k is a feasible solution of the QP subproblem if $\zeta(x_k, w_k) > 0$, hence, $[\psi(x_k, w_k)]_+ \nabla_x \psi(x_k, w_k)^\top d_k \leq -\alpha_k [\psi(x_k, w_k)]_+ \rho(x_k, w_k) = -\alpha_k \zeta(x_k, w_k)$, otherwise the inequality holds trivially. Moreover, the last inequality follows from Assumption 2.1-(ii) and Lemma A.3 and one can easily verify that $\|d_k\| \leq 2C_\phi + \alpha_0$ and from Assumption 2.2 we have $\|\nabla_x \psi(x_k, w_k)\| \leq C_\psi$. Therefore, combining (22) with (21), we obtain

$$\begin{aligned}
p(x_{k+1}) - p(x_k) &\leq -2\gamma_k \alpha_k \zeta(x_k, w_k) + 2\gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) \\
&\quad + 2\gamma_k L_{xy}^\psi [g(x_k)]_+ \|w_k - w^*(x_k)\| \|d_k\| + \frac{\gamma_k^2 L_p(x_k)}{2} \|d_k\|^2 \\
&\leq -2\gamma_k \alpha_k \zeta(x_k, w_k) + 2\gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) \\
&\quad + L_{xw}^\psi p(x_k) \|w_k - w^*(x_k)\|^2 + \frac{\gamma_k^2 (L_p(x_k) + 2L_{xw}^\psi)}{2} \|d_k\|^2.
\end{aligned}$$

Next, rearranging the above inequality, dividing both sides by 2, lead to the desired result. \square

A.4 Proof of Theorems 4.1 and 4.2

Before proving Theorems 4.1 and 4.2, we present a technical lemma on the recursive relation of a non-negative real-valued sequence that will be used in our convergence analysis.

Lemma A.6 ([3] Lemma 5.31). *Let $\{v_k\}$, $\{u_k\}$, $\{\alpha_k\}$, $\{\beta_k\}$ be sequences of nonnegative reals with $\sum_{k=0}^\infty \alpha_k < \infty$ and $\sum_{k=0}^\infty \beta_k < \infty$ such that $v_{k+1} \leq (1 + \alpha_k)v_k - u_k + \beta_k$ for all k . Then, $\{v_k\}$ converges and $\sum_{k=0}^\infty u_k < \infty$.*

Using this result, we first show that the sequence $\{\gamma_k \|d_k\|^2\}_k$ is summable and $\{p(x_k)\}_k$ is a bounded sequence.

Lemma A.7. *Let $\{x_k\}_k$ be the sequence generated by Algorithm 1 such that $\sum_{k=0}^{+\infty} \gamma_k \alpha_k < +\infty$, $\sum_{k=0}^{+\infty} \delta_y^{N_k} < +\infty$, and $\sum_{k=0}^{+\infty} \delta_w^{M_k} < +\infty$. Under the premises of Lemma A.5, we have that (i) $\sum_{k=0}^{+\infty} \gamma_k \|d_k\|^2 < +\infty$; (ii) $\{p(x_k)\}_{k \geq 0}$ is a bounded sequence, i.e., there exists $C_g > 0$ such that $[g(x_k)]_+ \leq C_g$ for any $k \geq 0$.*

Proof. (i) Consider **Part (I)** of Lemma A.5 by rearranging terms one can obtain:

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma_k}{2} \|d_k\|^2 + \gamma_k \alpha_k (C_\phi + \alpha_k) + \frac{L_{xy}^\phi}{2} \Delta_1^y \delta_y^{N_k}.$$

Since α_k is a non-increasing sequence and it is assumed that $\sum_{k=0}^{+\infty} \gamma_k \alpha_k < +\infty$, one can verify that $\sum_{k=0}^{+\infty} \gamma_k \alpha_k^2 \leq \sum_{k=0}^{+\infty} \gamma_k \alpha_k < +\infty$. Moreover, since $\sum_{k=0}^{+\infty} \delta_y^{N_k} < +\infty$, we have $\sum_{k=0}^{+\infty} (\gamma_k \alpha_k (C_\phi + \alpha_k) + \frac{L_{xy}^\phi}{2} \Delta_1^y \delta_y^{N_k}) < +\infty$. Therefore, applying Lemma A.6, we conclude that $\sum_{k=0}^{+\infty} \gamma_k \|d_k\|^2 < +\infty$.

(ii) Similarly, from **Part (II)** of Lemma A.5, multiplying both sides by 2, using $L_p(x) = 2C_\psi^2 + L_g^2 + p(x)$ from Lemma A.4, and rearranging terms yields:

$$\begin{aligned}
p(x_{k+1}) &\leq \underbrace{(1 + L_{xw}^\psi \Delta_1^w \delta_w^{M_k} + \frac{\gamma_k^2}{2} \|d_k\|^2) p(x_k)}_{\mathbf{a}_k} - 2\gamma_k \alpha_k \zeta(x_k, w_k) \\
&\quad + \underbrace{2\gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) + \frac{\gamma_k^2 (2C_\psi^2 + L_g^2 + 2L_{xw}^\psi)}{2} \|d_k\|^2}_{\mathbf{b}_k}.
\end{aligned}$$

From the assumptions in the statement of the lemma and the result of part (I) and that $\gamma_k \in (0, 1)$, we have that $\sum_{k=0}^{+\infty} \mathbf{a}_k < +\infty$ and $\sum_{k=0}^{+\infty} \mathbf{b}_k < +\infty$. Hence, the conditions of Lemma A.6 are satisfied, and we conclude that the sequence $\{p(x_k)\}_{k \geq 0}$ converges. Therefore, $\{p(x_k)\}_{k \geq 0}$ is bounded, i.e., there exists $C_g > 0$ such that $[g(x_k)]_+ \leq C_g$ for all $k \geq 0$. \square

Now, we are ready to prove Theorem 4.1. First, we restate the statement with full details here.

Theorem A.8 (Restatement of Theorem 4.1). *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Let $\{x_k, \lambda_k\}_{k \geq 0}$ be the sequence generated by Algorithm 1 such that $\{\alpha_k\}_k$ is a non-increasing sequence and $\gamma_k \leq (L_f + L_{xy}^\phi)^{-1}$. Then, for any $T \geq 1$ and $k \geq 1$,*

$$(I) \quad \frac{1}{\Gamma_T} \sum_{k=0}^{T-1} \gamma_k \|d_k\|^2 \leq \frac{2(f(x_0) - f(x_T))}{\Gamma_T} + \frac{1}{\Gamma_T} \sum_{k=0}^{T-1} \gamma_k \alpha_k (C_\phi + \alpha_k) + \frac{L_{xy}^\phi \Delta_1^y}{\Gamma_T} \sum_{k=0}^{T-1} \delta_y^{N_k}, \quad (23)$$

$$(II) \quad \frac{1}{A_T} \sum_{k=0}^{T-1} \alpha_k [g(x_k)]_+^{2\theta} \leq \frac{\mu}{A_T} \sum_{k=0}^{T-1} \left(\frac{p(x_k)}{\gamma_k} - \frac{p(x_{k+1})}{\gamma_k} \right) + \frac{\mu}{A_T} \sum_{k=0}^{T-1} \left(\frac{\bar{\Delta}}{\gamma_k} \delta_w^{M_k} + \frac{2\alpha_k}{\mu} (\Delta_2^w)^{2\theta} \delta_w^{2\theta M_k} \right) \\ + \frac{\mu(L_p + 2L_{xw}^\psi)}{2A_T} \sum_{k=0}^{T-1} \gamma_k \|d_k\|^2, \quad (24)$$

for some $\bar{\Delta} > 0$, where $\Gamma_T \triangleq \sum_{k=0}^{T-1} \gamma_k$ and $A_T \triangleq \sum_{k=0}^{T-1} \alpha_k$.

Proof. Part (I) follows immediately from Lemma A.5-Part (I) by summing over $k = 0$ to $T - 1$ and dividing both sides by $\Gamma_k = \sum_{k=0}^{T-1} \gamma_k$.

To prove Part (II), first note that from Lemma A.7 we have $[g(x_k)]_+ \leq C_g$ which from Lemma A.4 we conclude that there exists a constant $L_p \triangleq 2C_\psi^2 + L_g^2 + C_g^2$ that upper bounds the local Lipschitz constant $L_p(x)$ uniformly along the sequence $\{x_k\}_{k \geq 0}$. Therefore, we can simplify the bound in (19) as follows

$$\gamma_k \alpha_k \zeta(x_k, w_k) \leq \frac{p(x_k)}{2} - \frac{p(x_{k+1})}{2} + \gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) \\ + \frac{L_{xw}^\psi}{2} C_g^2 \Delta_1^w \delta_w^{M_k} + \frac{\gamma_k^2 (L_p + 2L_{xw}^\psi)}{4} \|d_k\|^2.$$

Using Assumption 2.3, we can lower bound the left-hand side of the above inequality by $\frac{\gamma_k \alpha_k}{\mu} [\psi(x_k, w_k)]_+^{2\theta}$. Moreover, from (17) and that $\theta \in (0, 1)$ we have that $\frac{1}{2} [g(x_k)]_+^{2\theta} \leq [\psi(x_k, w_k)]_+^{2\theta} + (\Delta_2^w)^{2\theta} \delta_w^{2\theta M_k}$ which leads to

$$\frac{\gamma_k \alpha_k}{2\mu} [g(x_k)]_+^{2\theta} \leq \frac{p(x_k)}{2} - \frac{p(x_{k+1})}{2} + \gamma_k \Delta_1^w \delta_w^{M_k} C_\psi (2C_\phi + \alpha_0) + \frac{L_{xw}^\psi}{2} C_g^2 \Delta_1^w \delta_w^{M_k} \\ + \frac{\gamma_k^2 (L_p + 2L_{xw}^\psi)}{4} \|d_k\|^2 + \frac{\gamma_k \alpha_k}{\mu} (\Delta_2^w)^{2\theta} \delta_w^{2\theta M_k}.$$

Finally, multiplying both sides by $2\mu/\gamma_k$, summing over $k = 0$ to $T - 1$, dividing by A_T , and defining $\bar{\Delta} \triangleq \Delta_1^w C_\psi (2C_\phi + \alpha_0)$ lead to the desired result. \square

Now, we restate and prove Theorem 4.2.

Theorem A.9 (Restatement of Theorem 4.2). *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Let $\{x_k, \lambda_k\}_{k \geq 0}$ be the sequence generated by Algorithm 1 such that for any $k \geq 0$, $\alpha_k = \frac{T^{1/3}}{(k+2)^{1+\omega}}$, $\gamma_k = \gamma = \min\{\frac{\mu C_g^{2-2\theta}}{T^{1/3}}, (L_f + L_{xy}^\phi)^{-1}\}$, $N_k = \frac{2}{1-\delta_y} \log(k+1)$, and $M_k = \frac{1}{1-\delta_w} \max\{\max\{1, \frac{1}{2\theta}\} \log(T), \log(T[\psi(x_k, w_k)]_+^{4\theta-2})\}$ if $[\psi(x_k, w_k)]_+ \|\nabla_x \psi(x_k, w_k)\| > 0$, otherwise, $M_k = \frac{1}{1-\delta_w} \max\{1, \frac{1}{2\theta}\} \log(T)$. Then, for any $\epsilon > 0$, there exists $t \in \{0, \dots, T-1\}$ such that*

$$1. \text{ (Stationarity) } \|\nabla f(x_t) + \lambda_t \nabla g(x_t)\| \leq \epsilon \text{ within } T = \mathcal{O}(\frac{1}{\epsilon^3}) \text{ iterations;}$$

2. (Feasibility) $[g(x_t)]_+ \leq \epsilon$ within $T = \mathcal{O}(\frac{1}{\epsilon^{6\theta}})$ iterations;
3. (Slackness) $|\lambda_t g(x_t)| \leq \epsilon$ within $T = \mathcal{O}(\frac{1}{\epsilon^{3\theta/(1-\theta)}})$ iterations.

Proof. Before starting the proof let us define $t \triangleq \operatorname{argmin}_{0 \leq k \leq T-1} \max\{\|\nabla f(x_k) + \lambda_k \nabla g(x_k)\|, [g(x_k)]_+, |\lambda_k g(x_k)|\}$. Moreover, the selection of parameters α_k , γ_k , N_k , and M_k implies that the conditions of Lemma A.7 hold, and we can invoke its result within the proof.

Part 1. First, we show the result for ϵ -stationary condition. From the definition of d_k and comparing it with $\nabla f(x_k) + \lambda_k \nabla g(x_k)$ we observe that if $\zeta(x_k, w_k) = 0$ then $\lambda_k = 0$ and $\|\nabla f(x_k) + \lambda_k \nabla g(x_k)\|^2 \leq 2\|d_k\|^2 + 2(L_{xy}^\phi \|y_k - y^*(x_k)\|)^2 \leq 2\|d_k\|^2 + 2(L_{xy}^\phi)^2 \Delta_1^y \delta_y^{N_k}$ which by selecting N_k as in the statement of corollary, we obtain $\|\nabla f(x_k) + \lambda_k \nabla g(x_k)\|^2 \leq 2\|d_k\|^2 + 2(L_{xy}^\phi)^2 \Delta_1^y \frac{1}{(k+1)^2}$. If $\zeta(x_k, w_k) > 0$, then

$$\begin{aligned}
& \|\nabla f(x_k) + \lambda_k \nabla g(x_k)\|^2 \\
& \leq 3\|d_k\|^2 + 3\|\nabla f(x_k) - \nabla_x \phi(x_k, y_k)\|^2 + 3\lambda_k^2 \|\nabla g(x_k) - \nabla_x \psi(x_k, w_k)\|^2 \\
& \leq 3\|d_k\|^2 + 3(L_{xy}^\phi)^2 \Delta_1^y \delta_y^{N_k} + 3\lambda_k^2 (L_{xw}^\psi)^2 \Delta_1^w \delta_w^{M_k} \\
& \leq 3\|d_k\|^2 + 3(L_{xy}^\phi)^2 \Delta_1^y \delta_y^{N_k} + 3\mu^2 C^2 [\psi(x_k, w_k)]_+^{2-4\theta} (L_{xw}^\psi)^2 \Delta_1^w \delta_w^{M_k} \\
& \leq 3\|d_k\|^2 + 3(L_{xy}^\phi)^2 \Delta_1^y \frac{1}{(k+1)^2} + 3\mu^2 C^2 (L_{xw}^\psi)^2 \Delta_1^w \frac{1}{T}, \tag{25}
\end{aligned}$$

where in the second inequality we used Lipschitz continuity of $\nabla_x \phi$ and $\nabla_x \psi$ as well as the relations in (15) and (16). The third inequality follows from Lemma A.3 and Assumption 2.3 which shows that $\lambda_k \leq C\|\nabla_x \psi(x_k, w_k)\|^{-1} \leq C\mu[\psi(x_k, w_k)]_+^{1-2\theta}$ for some $C > 0$. The last inequality is obtain by plugging the selection of N_k and M_k as in the statement of corollary and noting that $\frac{1}{1-\delta} \geq 1/\log(1/\delta)$ for any $\delta \in (0, 1)$.

On the other hand, from Theorem A.8 part (I), by selecting $\gamma = \mathcal{O}(1/T^{1/3})$ and $\alpha_k = \frac{T^{1/3}}{(k+2)^{1+\omega}}$ and noting that $\frac{1}{T} \sum_{k=0}^{T-1} \alpha_k = \mathcal{O}(1/T^{2/3})$, we conclude that $\frac{1}{T} \sum_{k=0}^{K-1} \|d_k\|^2 \leq \mathcal{O}(1/T^{2/3})$. Therefore, combining the result with (25) we obtain

$$\|\nabla f(x_t) + \lambda_t \nabla g(x_t)\|^2 \leq \frac{1}{T} \sum_{k=0}^{K-1} \|\nabla f(x_k) + \lambda_k \nabla g(x_k)\|^2 \leq \mathcal{O}\left(\frac{1}{T^{2/3}} + \frac{(L_{xw}^\psi)^2 \Delta_1^w}{T}\right).$$

By taking the square root of both sides of the above inequality, the result of part 1 follows immediately.

Part 2. From Lemma A.7, we observe that there exists $D > 0$ such that $D = \sum_{k=0}^{T-1} \gamma \|d_k\|^2 < +\infty$. Considering the result of Theorem A.8-part (II), selecting $\gamma_k = \gamma = \mathcal{O}(\frac{1}{T^{1/3}})$, and $p(x) \geq 0$, we have that

$$\begin{aligned}
\frac{1}{A_T} \sum_{k=0}^{T-1} \alpha_k [g(x_k)]_+^{2\theta} & \leq \frac{\mu}{A_T \gamma} p(x_0) + \frac{\mu}{A_T} \sum_{k=0}^{T-1} \left(\frac{\bar{\Delta}}{\gamma} \delta_w^{M_k} + \frac{2\alpha_k}{\mu} (\Delta_2^w)^{2\theta} \delta_w^{2\theta M_k} \right) + \frac{\mu(L_p + 2L_{xw}^\psi)}{2A_T} D \\
& \leq \mathcal{O}\left(\frac{1}{A_T \gamma} + \frac{D}{A_T}\right),
\end{aligned}$$

where the last inequality follows from plugging in M_k since $\max\{\delta_w^{M_k}, \delta_w^{2\theta M_k}\} = \mathcal{O}(\frac{1}{T})$. Therefore, from the above inequality, noting that $A_T = \Omega(T^{1/3})$, and the definition of t at the beginning of the proof we conclude that $[g(x_t)]_+^{2\theta} \leq \mathcal{O}(\frac{1}{T^{1/3}})$ which completes the proof of part 2.

Part 3. Finally, to calculate the complexity of finding ϵ -complementarity slackness, recall the update of λ_k in Algorithm 1. Recall that $\zeta(x_k, w_k)[\psi(x_k, w_k)]_+ + \|\nabla_x \psi(x_k, w_k)\|$. If $\zeta(x_k, w_k) = 0$, then $\lambda_k = 0$, hence, $\lambda_k g(x_k) = 0$. Suppose $\zeta(x_k, w_k) > 0$, then we observe that $g(x_k) \geq \psi(x_k, w_k) > 0$. Therefore, from Lemma A.3 we have that $0 \leq \lambda_k g(x_k) = \lambda_k [g(x_k)]_+ \leq C[g(x_k)]_+^{2-2\theta} + C\Delta_2^w \delta_w^{M_k} [\psi(x_k, w_k)]_+^{1-2\theta}$. Combining the two scenarios, for any $k \geq 0$, we have that $|\lambda_k g(x_k)| \leq C[g(x_k)]_+^{2-2\theta} + C\Delta_2^w \delta_w^{M_k} [\psi(x_k, w_k)]_+^{1-2\theta}$ for some $C > 0$. Therefore, based on selection of M_k , we obtain $|\lambda_t g(x_t)| \leq \mathcal{O}(\frac{1}{T^{(1-\theta)/(3\theta)}} + \frac{1}{T})$ from which the result follows. \square

A.5 Experiment Details and Additional Plots

Experiment Details: In all experiments, we select the regularization parameter $\lambda = 10^{-3}$ and the maximization variables y, w are updated by running $N_k = 2\lceil\log(k+2)\rceil$ and $M_k = 10\lceil\log(k+2)\rceil$ steps of the projected gradient ascent method. The stepsize γ is tuned by selecting the best performance among $\{10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ and the parameter is set $\alpha_k = \alpha/(k+2)^{1.001}$ for $\alpha \in \{0.1, 0.2, 0.5, 1\}$. Hyperparameter choices follow Theorem 4.2 and tuned via targeted grid search to ensure robustness. Furthermore, to determine the threshold value r , we solve the robust learning task in the constraint, i.e., $\min_x \max_{w \in \Delta_m} \sum_{j=1}^m \ell_2(x, \xi_j^{(2)}) - g_m(w)$, separately using the unconstrained variant of our method for a some iterations. The resulting objective value is then used in the original problem as the threshold value.

The oscillations that occur in plots reflect the difficult trade-off between minimizing the objective, enforcing feasibility under infinitely many functional constraints, and satisfying the ϵ -KKT conditions, a behavior common in both convex and nonconvex problems with functional constraints [27].

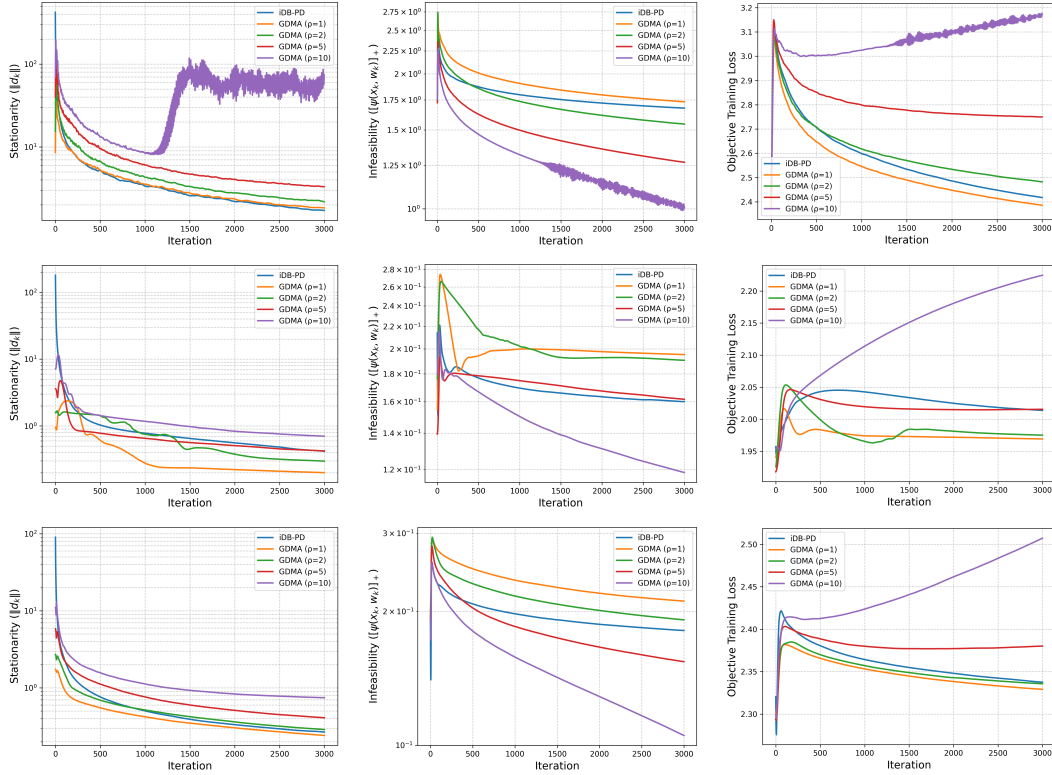


Figure 3: iDB-PD vs. GDMA on multi-Fashion MNIST (top row), Yeast (middle row), and 20NG (bottom row), evaluated in terms of stationarity, infeasibility, and objective loss.

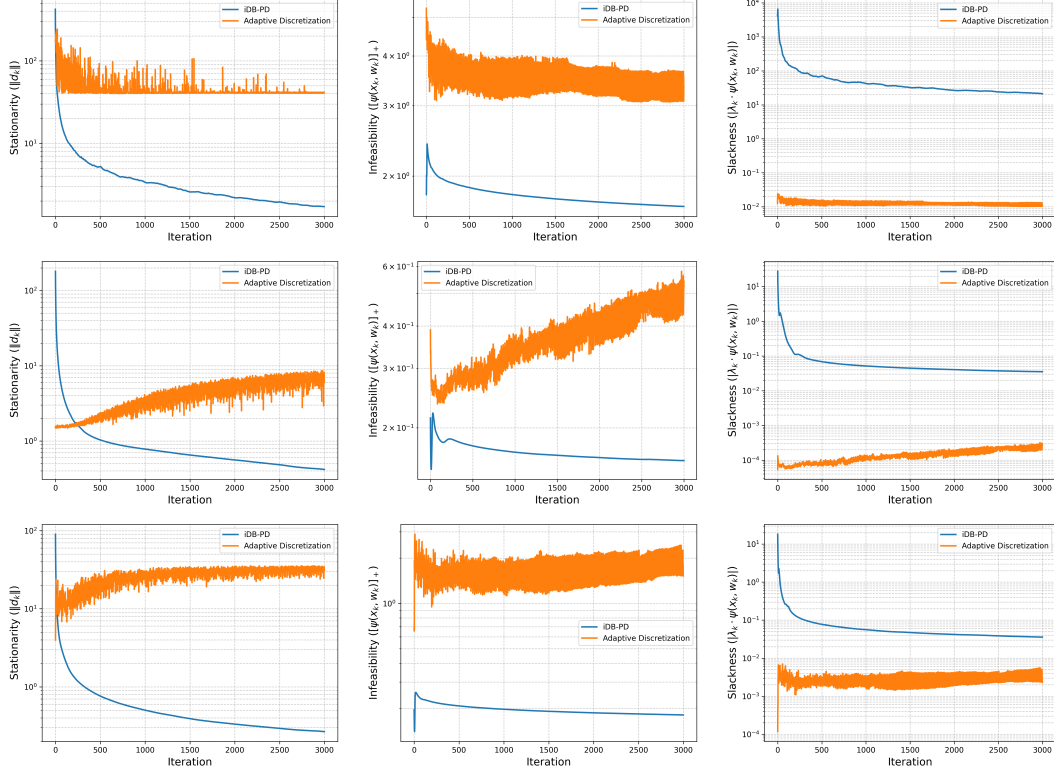


Figure 4: iDB-PD vs. Adaptive Discretization with COOPER on multi-Fashion MNIST (top row), Yeast (middle row), and 20NG (bottom row), evaluated in terms of stationarity, infeasibility, and slackness.

Across the three additional datasets, iDB-PD broadly outperforms all GDMA variants and the adaptive discretization method with COOPER. iDB-PD drives infeasibility and stationarity down quickly while maintaining competitive objective values. In contrast, GDMA requires large penalty values to approach feasibility, frequently at the cost of stability. Further, adaptive discretization struggles with instability and struggles with matching iDB-PD’s stationarity and infeasibility performance. These results confirm the robustness of our iDB-PD method which effectively balances feasibility, optimality, and stability.