SMOOTH MINIMAX OPTIMISATION IN NON-EUCLIDEAN SPACE USING BREGMAN DIVERGENCE FRAMEWORK

Sandeep K. Routray
170623
sroutray@iitk.ac.in

Nitish V. Deshpande
170450
nitishvd@iitk.ac.in

ABSTRACT

Minimax type of problems arise in several domains such as machine learning, optimization, statistics, communication, and game theory. However, a majority of results are established for the Euclidean norm due to its special self-dual nature. In this project, we propose Generalized Conceptual Dual Implicit Accelerated Gradient Descent (GC-DIAG) which is adapted from the Conceptual Dual Implicit Accelerated Gradient (C-DIAG) [9] for solving smooth minimax optimization problems $\min_{\mathbf{x}} \max_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})$ where $g(.,.)$ is smooth and $g(\mathbf{x}, .)$ is concave for each $\mathbf{x}$. We prove $\mathcal{O}(1/k^2)$ convergence rate for the primal dual gap using a potential-function based proof [2] and Bregman Divergence framework. We also prove $\mathcal{O}(\frac{1}{k^4})$ convergence rate using Nesterov's accelerated gradient descen$\hat{a}$nd a restarting strategy, similar to [10], which is an improvement over $\mathcal{O}(\frac{1}{k})$ for smooth and strongly convex functions with respect to an arbitrary norm.

1 Motivation

The motivation for the non-Euclidean extension comes from the lecture notes on mirror descent in [4] which mentions that the quadratic proximity term $\frac{1}{2\eta_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2$ is inappropriate for problems with highly inhomogeneous geometry. For example, the quadratic minimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{Q} (\mathbf{x} - \mathbf{x}_0)$ where $\mathbf{Q} \succ \mathbf{0}$ is a diagonal matrix with high condition number. It is also possible to have non-euclidean geometry, for instance, minimization with constraint set as the probability simplex. The inhomogeneous geometry leads to slow updates (since iteration complexity depends on condition number for the quadratic minimization problem) in the conventional GD algorithm. For the probability simplex problem, Euclidean distance is in general not recommended for measuring the distance between probability vectors. To tackle these issues, the mirror gradient descent algorithm was introduced which adjusts the gradient updates to fit the problem geometry. The notion of mirror GD is to replace the quadratic proximity term $\frac{1}{2\eta_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2$ by a class of general distance-like metric known as the Bregman divergence [4].

2 Prior work

Thekumparampil et al [9] proposed a new algorithm combining Nesterov's AGD and Mirror-Prox and proved a convergence rate $\tilde{\mathcal{O}}(1/k^2)$ for smooth, strongly-convex – concave problems, improving upon the previous best known rate of $\mathcal{O}(1/k)$. However, their analysis is limited to the Euclidean norm. For extending an algorithm to the Non-Euclidean space, there have been two popular approaches. First, algorithms can be analyzed in vector spaces defined with respect to an arbitrary norm. Bansal et al [2] analyzed Nesterov's accelerated gradient descent using general norm for the unconstrained case. They develop a potential function based framework for proving convergence rate. A second and more recent approach is the notion of relative strong convexity and smoothness [5]. Haihao Lu et al in [5] develop a notion of "relative smoothness" and relative strong convexity that is determined relative to a user-specified "reference function" $h(.)$. However, extension of this notion to accelerated gradient descent is still an open problem. Since, analysis of extension to Non-Euclidean space is tough, some works analyze algorithms in a special class of Non-Euclidean spaces called as the Riemannian space. Kwangjun Ahn in [1] proposed a Riemannian counterpart to Nesterov's AGD.

3 Contributions

(1) We develop a potential function based proof of $\mathcal{O}(1/k^2)$ convergence rate for Nesterov's AGD for the *general norm* and *constrained case* using Bregman divergence framework.

(2) We propose Generalized Conceptual Dual Implicit Accelerated Gradient Descent (GC-DIAG) which is adapted from the Conceptual Dual Implicit Accelerated Gradient (C-DIAG) [9] and proved $\mathcal{O}(1/k^2)$ convergence rate for the primal dual gap. Our attempt to develop an implementable inexact version of GC-DIAG was unsuccessful.

(3) We propose a restarting strategy using Nesterov's AGD for smooth and strongly convex case w.r.t arbitrary norm and show that the convergence rate can be improved from $\mathcal{O}(\frac{1}{k^2})$ to $\mathcal{O}(\frac{1}{k^4})$. These results along with similar results for mirror descent are summarized in Table 1.

	Smooth and convex	Smooth and strongly convex
Mirror descent	$\mathcal{O}(\frac{1}{\sqrt{k}})$	$\mathcal{O}(\frac{1}{k})$ [10]
Nesterov's AGD	$\mathcal{O}(\frac{1}{k^2})$ [7, 2]	$\mathcal{O}(\frac{1}{k^4})$

Table 1: Comparison of oracle complexities with arbitrary norm

4 Problem Formulation

We consider smooth minimax problems of the form

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x},\mathbf{y}) \quad , \quad g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \tag{1}$$

under the following assumptions

(A1) $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is L-smooth *i.e.* , gradient Lipschitz (see Definition 1).

(A2) $g(\mathbf{x}, .)$ is concave for every $\mathbf{x}$ and σ-strongly convex for every $g(., \mathbf{y})$ for every $\mathbf{y}$.

(A3) $\mathcal{Y}$ is a compact set *i.e.* , there exists a finite $D_{\mathcal{Y}} = \max_{\mathbf{y},\mathbf{y}'\in\mathcal{Y}} \|\mathbf{y}-\mathbf{y}'\|$ also known as the diameter of $\mathcal{Y}$.

(A4) There exists a finite $\Omega = \max_{\mathbf{y}\in\mathcal{Y}} D_\psi(\mathbf{y}||\mathbf{y}_0)$.

Note that $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}$ where $\mathcal{V}$ is a normed vector space with an arbitrary norm $\|.\|$ on the underlying space. We have marked in red the parts where our work differs from the original paper [9].

5 Preliminaries

Definition 1. *A function* $g(\mathbf{x},\mathbf{y})$ *is said to be* L*-smooth if:*

$$\max\left\{\|\nabla_{\mathbf{x}} g(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}',\mathbf{y}')\|_*, \|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} g(\mathbf{x}',\mathbf{y}')\|_*\right\} \le L\left(\|\mathbf{x}-\mathbf{x}'\| + \|\mathbf{y}-\mathbf{y}'\|\right)$$

5.1 Bregman divergences

Definition 2. Given a continuously differentiable and μ_ψ-strongly convex function $\psi : \mathcal{X} \to \mathbb{R}$, the Bregman divergence is defined as

$$D_\psi(\mathbf{y}\|\mathbf{x}) := \psi(\mathbf{y}) - \psi(\mathbf{x}) - \langle\nabla\psi(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \tag{2}$$

The strong convexity of ψ implies

$$D_\psi(\mathbf{y}\|\mathbf{x}) \ge \frac{\mu_\psi}{2}\|\mathbf{y}-\mathbf{x}\|^2 \tag{3}$$

5.2 Convex-concave setting

The convex concave minimax Problem 1 induces the following primal and dual problems

$$P^* = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \quad f(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x},\mathbf{y}) \quad (P) \tag{4}$$

$$D^* = \max_{\mathbf{y} in \mathcal{Y}} h(\mathbf{y}), \quad h(\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}) \quad (D) \tag{5}$$

Under Assumption (**A2**), $\forall(\hat{\mathbf{x}},\hat{\mathbf{y}})$, the following holds trivially:

$$\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\hat{\mathbf{y}}) \le g(\hat{\mathbf{x}},\hat{\mathbf{y}}) \le \max_{\mathbf{y}\in\mathcal{Y}} g(\hat{\mathbf{x}},\mathbf{y})$$

which then implies $\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}) \le \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x},\mathbf{y})$. Further, under Assumption (**A3**), Sion's minimax theorem [8] states the above inequality is in fact a equality *i.e.* , $\max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x},\mathbf{y})$. Thus, any point $(\mathbf{x}^*,\mathbf{y}^*)$ is an optimal solution to Problem 1 if and only if

$$D^* = h(\mathbf{y}^*) = \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}^*) = g(\mathbf{x}^*,\mathbf{y}^*) = \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x}^*,\mathbf{y}) = f(\mathbf{x}^*) = P^* \tag{6}$$

i.e. $\mathbf{x}^*$ is an optimal solution to (P) and $\mathbf{y}^*$ is an optimal solution to (D).

We would like to find a ϵ-primal-dual pair $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ defined as:

Definition 3. *For a convex-concave $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is an ϵ-primal-dual pair of g if the primal-dual gap is less than ϵ: $\max_{\mathbf{y}\in\mathcal{Y}} g(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\hat{\mathbf{y}}) = f(\hat{\mathbf{x}}) - h(\hat{\mathbf{y}}) \le \epsilon$*

Note that the ϵ-primal-dual criteria for $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ can also be written as

$$f(\hat{\mathbf{x}}) - P^* + D^* - h(\hat{\mathbf{y}}) \le \epsilon$$

which implies that

$$f(\hat{\mathbf{x}}) - P^* \le \epsilon \tag{7}$$

$$D^* - h(\hat{\mathbf{y}}) \le \epsilon \tag{8}$$

i.e. $\hat{\mathbf{x}}$ is an ϵ-optimal minima of f and $\hat{\mathbf{y}}$ is an ϵ-optimal maxima of h.

5.3 Motivating the algorithm

Lemma 1 guarantees that the dual function $h(\mathbf{y})$ is an $L(1+\frac{L}{\sigma})$-smooth concave function. So, we can use AGD to ensure that the dual gap $h(\mathbf{y}_k) - h(\mathbf{y}^*) = \mathcal{O}(1/k^2)$. Each step of AGD updates for $\mathbf{y}$ requires gradients at $g(\mathbf{x}_{k+1},\mathbf{w}_k)$ which is computed by $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_k)$. In ℓ_2 norm space, $\mathbf{x}_{k+1}$ can be computed efficiently in logarithmic number of steps because $g(.,\mathbf{y}_k)$ is smooth and strongly-convex. So, the overall oracle complexity for dual problem is $h(\mathbf{y}_k) - h(\mathbf{y}^*) = \tilde{\mathcal{O}}(1/k^2)$ in ℓ_2 norm space. It is clear that in arbitrary norm space, the oracle complexity would be **worse** than $\tilde{\mathcal{O}}(1/k^2)$ as AGD does not achieve linear convergence for strong convex case for general norm.

Further, the above bound on the dual gap need not hold for primal gap as well. [9, Section 3] provides an example where the dual gap bound is $\Theta(1/k^2)$ but the primal gap bound is $\Theta(1/k)$ only. Equations (7), (8) imply that it is necessary to ensure the primal gap has same or better convergence rate than dual gap, so that the overall convergence rate is dictated by the dual problem. DIAG achieves this by combining ideas from AGD and Nemirovski's derivation of the Mirror-Prox algorithm [6].

Lemma 1. *For a σ-strongly-convex-concave L-smooth function $g(.,.)$, $h(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y})$ is an $L(1+\frac{L}{\sigma})$-smooth concave function.*

Proof. Let $\mathbf{x}^*(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y})$. Since $g(.,\mathbf{y})$ is strongly convex, $\mathbf{x}^*(\mathbf{y})$ is unique. Then by Danskin's theorem [3, Section 6.11], h is differentiable and $\nabla h(\mathbf{y}) = \nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}),\mathbf{y})$. First, we show that $\mathbf{x}^*(\mathbf{y})$ is $\frac{L}{\sigma}$-Lipschitz continuous as follows

$$\begin{aligned}
\sigma \left\|\mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_2)\right\|^2 &\overset{(a)}{\le} \langle \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_2),\mathbf{y}_2) - \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1),\mathbf{y}_2), \mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_1)\rangle \\
&\overset{(b)}{\le} \langle -\nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1),\mathbf{y}_2), \mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_1)\rangle \\
&\overset{(c)}{\le} \langle \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1),\mathbf{y}_1) - \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1),\mathbf{y}_2), \mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_1)\rangle
\end{aligned}$$

$$\begin{aligned}&\overset{(d)}{\leq} \|\nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1), \mathbf{y}_1) - \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}_1), \mathbf{y}_2)\|_* \|\mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_1)\| \\ &\overset{(e)}{\leq} L \|\mathbf{y}_1 - \mathbf{y}_2\| \|\mathbf{x}^*(\mathbf{y}_2) - \mathbf{x}^*(\mathbf{y}_1)\| \end{aligned} \tag{9}$$

where (a) follows from σ-strong convexity of $g(., y)$, (b) and (c) follows from the first order optimality condition for $\mathbf{x}^*(\mathbf{y})$: $\langle \nabla_{\mathbf{x}} g(\mathbf{x}^*(\mathbf{y}), \mathbf{y}), \mathbf{x} - \mathbf{x}^*(\mathbf{y}) \rangle \geq 0$ applied at $\mathbf{y} = \mathbf{y}_2$ and $\mathbf{y} = \mathbf{y}_1$ respectively, (d) follows from the Generalized Cauchy-Schwarz inequality and (e) follows from L-smoothness of g (Definition 1).

Now, to show h is smooth, we proceed as follows:

$$\begin{aligned}\|\nabla h(\mathbf{y}_1) - \nabla h(\mathbf{y}_2)\|_* &= \|\nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}_1), \mathbf{y}_1) - \nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}_2), \mathbf{y}_2)\|_* \\ &\overset{(a)}{\leq} L \|\mathbf{y}_1 - \mathbf{y}_2\| + L \|\mathbf{x}^*(\mathbf{y}_1) - \mathbf{x}^*(\mathbf{y}_2)\| \\ &\overset{(b)}{\leq} L \|\mathbf{y}_1 - \mathbf{y}_2\| + \frac{L}{\sigma} \cdot L \|\mathbf{y}_1 - \mathbf{y}_2\| = L(1 + \frac{L}{\sigma}) \|\mathbf{y}_1 - \mathbf{y}_2\| \end{aligned} \tag{10}$$

Here, (a) follows from L-smoothness of g and (b) follows from $\frac{L}{\sigma}$-Lipschitz continuity of $\mathbf{x}^*(\mathbf{y})$ (Equation (9)) □

5.4 Mirror-Prox

Mirror-Prox [6] is one of the popular algorithms employed to solve convex-concave minimax problem. It achieves an bound of $\mathcal{O}(1/k)$ on both primal and dual error. The *conceptual* Mirror-Prox (CMP) proposed in [6] brings out the main idea behind its $\mathcal{O}(1/k)$ convergence and motivates the final algorithm. CMP does the following updates:

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = (\mathbf{x}_k, \mathbf{y}_k) + \frac{1}{\beta}(-\nabla_{\mathbf{x}} g(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \nabla_{\mathbf{y}} g(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \tag{11}$$

CMP differs from standard gradient descent ascent in the point at which gradients are computed,*i.e.* in the k-th step, CMP uses gradient information at $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ for its updates. But updates given by Equation (11) is not implementable. In [6], an *inexact* version of CMP is proposed which can be efficiently implemented for smooth function $g(., .)$. It can be summarised as follows:

Let $\left(\mathbf{x}_k^{(0)}, \mathbf{y}_k^{(0)}\right) = (\mathbf{x}_k, \mathbf{y}_k)$. For $\beta < \frac{1}{L}$, the iteration

$$\left(\mathbf{x}_k^{(i+1)}, \mathbf{y}_k^{(i+1)}\right) = (\mathbf{x}_k, \mathbf{y}_k) + \frac{1}{\beta}\left(-\nabla_{\mathbf{x}} g\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}\right), \nabla_{\mathbf{y}} g\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}\right)\right) \tag{12}$$

can be shown to be a $\frac{1}{\sqrt{2}}$-contraction with $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ as its fixed point. So, in $\log\left(\frac{1}{\epsilon}\right)$ iterations of (12), we can obtain an accurate version of the update required by CMP. The $\texttt{Imp} - \texttt{STEP}$ of GC-DIAG (Algorithm 3) is motivated from Mirror-prox. We attempt to prove its convergence similarly, by trying to show a contraction.

5.5 Nesterov's accelerated gradient ascent with general norm

Nesterov's accelerated gradient descent (or ascent) [7] is a popular method used to minimize smooth convex functions (or maximize smooth concave functions) and has been shown to be an optimal method for this class of problems. The pseudocode for the general norm case is presented in Algorithm 1.

Algorithm 1 Nesterov's accelerated gradient ascent with general norm

```
Input: Smooth concave function $h(.)$, learning rate $\frac{1}{\beta}$, Bregman divergence $D_\psi(.\|.)$, initial point $\mathbf{y}_0$ and $\mathbf{z}_0$
Output: $\mathbf{y}_K$
for $k = 0, 1, ..., K$ do
    $\mathbf{w}_k \leftarrow (1 - \tau_k)\mathbf{y}_k + \tau_k \mathbf{z}_k$    (13)
    $\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y} \in \mathcal{Y}} \left\{ - \langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{w}_k\|^2 \right\}$    (14)
    $\mathbf{z}_{k+1} \leftarrow \arg\min_{\mathbf{z} \in \mathcal{Y}} \{ -\eta_k \langle \nabla h(\mathbf{w}_k), \mathbf{z} \rangle + D_\psi(\mathbf{z}\|\mathbf{z}_k) \}$    (15)
end for
```

We would consider a potential function based proof of Algorithm 1. If y^* is the optimal solution of the problem, the potential function is defined as

$$\Phi(k) = k\,(k+1)\,(h(\mathbf{y}^*) - h(\mathbf{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}^* \| \mathbf{z}_k) \tag{16}$$

The following lemma is the key result is obtaining the convergence rate of GENERALIZED DIAG (G-DIAG). It is essentially a generalization of Lemma 4 in [9] to arbitrary norm space. We emphasize the key steps in the proof using colours blue and green to improve the clarity of the proof.

Lemma 2. *Suppose $h(.)$ is an L-smooth function and the parameters of Algorithm 1 are chosen so that $\beta > L$, $\eta_k = \frac{(k+1)}{2\beta}\mu_\psi$ and $\tau_k = \frac{2}{k+2}$. Then, we have*

$$\Phi(k+1) \le \Phi(k)$$

Proof. Using (16), the potential difference can be written as

$$\begin{aligned}\Phi(k+1) - \Phi(k) &= (k+1)(k+2)\underbrace{(h(\mathbf{w}_k) - h(\mathbf{y}_{k+1}))}_{(a)} \\ &\underbrace{-k(k+1)(h(\mathbf{w}_k) - h(\mathbf{y}_k)) + 2(k+1)(h(\mathbf{y}) - h(\mathbf{w}_k))}_{(b)} + \frac{4\beta}{\mu_\psi}\underbrace{(D_\psi(\mathbf{y}\|\mathbf{z}_{k+1}) - D_\psi(\mathbf{y}\|\mathbf{z}_k))}_{(c)}\end{aligned} \tag{17}$$

The term (c) can be bounded as

$$\begin{aligned}D_\psi(\mathbf{y}\|\mathbf{z}_{k+1}) - D_\psi(\mathbf{y}\|\mathbf{z}_k) &= (\psi(\mathbf{y}) - \psi(\mathbf{z}_{k+1}) - \langle \nabla\psi(\mathbf{z}_{k+1}), \mathbf{y} - \mathbf{z}_{k+1}\rangle) - (\psi(\mathbf{y}) - \psi(\mathbf{z}_k) - \langle \nabla\psi(\mathbf{z}_k), \mathbf{y} - \mathbf{z}_k\rangle) \\ &= \psi(\mathbf{z}_k) - \psi(\mathbf{z}_{k+1}) + \langle \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{z}_k\rangle + \langle \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle \\ &\le -\frac{\mu_\psi}{2}\left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 + \underbrace{\langle \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle}_{(d)}\end{aligned} \tag{18}$$

where the last inequality is due to μ_ψ-strongly convex function $\psi(.)$. From the update in (15) in Algorithm 1, we write the optimality condition as

$$\left\langle (-\eta_k \nabla h(\mathbf{w}_k) + \nabla_{\mathbf{z}} D_\psi(\mathbf{z}\|\mathbf{z}_k))\big|_{\mathbf{z}=\mathbf{z}_{k+1}}, \mathbf{y} - \mathbf{z}_{k+1}\right\rangle \ge 0, \quad \forall \mathbf{y} \in \mathcal{Y}. \tag{19}$$

From definition of Bregman divergence, $\nabla_{\mathbf{z}} D_\psi(\mathbf{z}\|\mathbf{z}_k)\big|_{\mathbf{z}=\mathbf{z}_{k+1}} = \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k)$. Hence, the term (d) in Equation (18) can be bounded as

$$\langle \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle \le \langle \eta_k \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle\,. \tag{20}$$

Hence,

$$D_\psi(\mathbf{y}\|\mathbf{z}_{k+1}) - D_\psi(\mathbf{y}\|\mathbf{z}_k) \le -\frac{\mu_\psi}{2}\left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 + \langle \eta_k \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle\,. \tag{21}$$

The terms (a) and (b) can be bounded similar to the proof in [9, Lemma 4]. We provide the proof for the bound on terms (a) and (b) here for sake of completeness.

The term (a) can be bounded as

$$h(\mathbf{y}_{k+1}) - h(\mathbf{w}_k) \overset{(1)}{\ge} \langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k\rangle - \frac{L}{2}\left\|\mathbf{y}_{k+1} - \mathbf{w}_k\right\|^2 \overset{(2)}{\ge} \langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k\rangle - \frac{\beta}{2}\left\|\mathbf{y}_{k+1} - \mathbf{w}_k\right\|^2 \tag{22}$$

Here, inequality (1) follows from the fact that $(-h(\mathbf{x}))$ is L-smooth, inequality (2) follows from the choice of $\beta > L$. From the update in (14) in Algorithm 1, $\mathbf{y}_{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \left\{\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k\rangle - \frac{\beta}{2}\left\|\mathbf{y} - \mathbf{w}_k\right\|^2\right\}$. We know that $\mathbf{y}_k \in \mathcal{Y}$ and $\mathbf{z}_{k+1} \in \mathcal{Y}$. So, a convex combination $\mathbf{v} = (1-\tau_k)\mathbf{y}_k + \tau_k \mathbf{z}_{k+1} \in \mathcal{Y}$. Hence, we can write

$$\langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k\rangle - \frac{\beta}{2}\left\|\mathbf{y}_{k+1} - \mathbf{w}_k\right\|^2 \ge \langle \nabla h(\mathbf{w}_k), \mathbf{v} - \mathbf{w}_k\rangle - \frac{\beta}{2}\left\|\mathbf{v} - \mathbf{w}_k\right\|^2$$

$$\overset{(3)}{=} \tau_k \left\langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{z}_k \right\rangle - \frac{\beta}{2}\tau_k^2 \left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2, \tag{23}$$

where equality (3) follows from the update rule (13) in Algorithm 1.

The term (b) can be bounded as

$$\begin{aligned}
&-k(k+1)(h(\mathbf{w}_k) - h(\mathbf{y}_k)) + 2(k+1)(h(\mathbf{y}) - h(\mathbf{w}_k)) \\
&\overset{(4)}{\leq} -k(k+1)\left\langle \nabla h(\mathbf{w}_k), \mathbf{w}_k - \mathbf{y}_k \right\rangle + 2(k+1)\left\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \right\rangle \overset{(5)}{=} 2(k+1)\left\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{z}_k \right\rangle,
\end{aligned} \tag{24}$$

where inequality (4) follows from the concavity of $h(\mathbf{x})$ and equality (5) follows from the update rule (13) in Algorithm 1 and the choice of $\tau_k = \frac{2}{k+2}$. We now substitute bounds (21), (72) and (73) in (17) to get

$$\begin{aligned}
\Phi(k+1) - \Phi(k) \leq{}& (k+1)(k+2)\left(-\tau_k \left\langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{z}_k \right\rangle + \frac{\beta}{2}\tau_k^2 \left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2\right) \\
&+ 2(k+1)\left\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{z}_k \right\rangle + \frac{4\beta}{\mu_\psi}\left(-\frac{\mu_\psi}{2}\left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 + \left\langle \eta_k \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y} \right\rangle\right) \\
&\overset{(6)}{\leq} 2\beta \left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 \left(\frac{k+1}{k+2} - 1\right) + (-2(k+1) + \frac{4\beta}{\mu_\psi}\eta_k) \left\langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y} \right\rangle \overset{(7)}{\leq} 0,
\end{aligned} \tag{25}$$

where inequality (6) follows from the choice of τ_k and inequality (7) follows from the choice of $\eta_k = \frac{(k+1)}{2\beta}\mu_\psi$. □

We also attempted along similar lines using the notion of relative smoothness. We were not able to complete the proof. We highlight the steps which cause difficulty in the analysis in the Appendix.

5.6 Experiments

We demonstrate the significance of considering an appropriate norm which fits the problem geometry in the following example.

AGD vs general AGD

We consider the following problem and solve it using Algorithm 1.

$$\max_{\mathbf{y}} -\frac{1}{2}\mathbf{y}^T\mathbf{Q}\mathbf{y} := h(\mathbf{y}), \quad \mathbf{Q} \succ \mathbf{0}, \mathbf{y} \in \mathbb{R}^n \tag{26}$$

Here, $\mathbf{y}^* = 0$ and $h(\mathbf{y}^*) = 0$. Consider two different scenarios:

Case1: We minimize Problem 26 in ℓ^2 norm space. With respect to this norm, $h(.)$ is $\lambda_{\max}(\mathbf{Q})$-smooth. The AGD case corresponds to setting $\psi(\mathbf{y}) = (1/2)\mathbf{y}^T\mathbf{y} = (1/2)\left\|\mathbf{y}\right\|_2^2$ which has $\mu_\psi = 1$. The Bregman divergence becomes $D_\psi(\mathbf{y}||\mathbf{x}) = (1/2)\left\|\mathbf{y} - \mathbf{x}\right\|_2^2$. Therefore, we get the following update equations:

$$\mathbf{w}_k = (1 - \tau_k)\mathbf{y}_k + \tau_k\mathbf{z}_k \tag{27}$$

$$\mathbf{y}_{k+1} = \mathbf{w}_k - \frac{1}{\beta}\mathbf{Q}\mathbf{w}_k \tag{28}$$

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta_k\mathbf{Q}\mathbf{w}_k \tag{29}$$

The variation of optimality gap with iterations is plotted in Figure 1.

Case2: We minimize Problem 26 in $\mathbf{Q}$-norm space, *i.e.* $\left\|\mathbf{y}\right\|_{\mathbf{Q}} = \sqrt{\mathbf{y}^T\mathbf{Q}\mathbf{y}}$. With respect to this norm, $h(.)$ is 1-smooth. We set $\psi(\mathbf{y}) = (1/2)\mathbf{y}^T\mathbf{Q}\mathbf{y} = (1/2)\left\|\mathbf{y}\right\|_{\mathbf{Q}}^2$ which has $\mu_\psi = 1$ with respect to $\mathbf{Q}$-norm. The Bregman divergence becomes $D_\psi(\mathbf{y}||\mathbf{x}) = (1/2)\left\|\mathbf{y} - \mathbf{x}\right\|_{\mathbf{Q}}^2$. Therefore, we get the following update equations:

$$\mathbf{w}_k = (1 - \tau_k)\mathbf{y}_k + \tau_k\mathbf{z}_k \tag{30}$$

$$\mathbf{y}_{k+1} = \mathbf{w}_k - \frac{1}{\beta}\mathbf{w}_k \tag{31}$$

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta_k\mathbf{w}_k \tag{32}$$

With $\beta = 1$, we have $\mathbf{y}_1 = \mathbf{0}$. So, we get the optimal solution with just one iteration of AGD. This shows that the $\mathbf{Q}$-norm space is a good choice of norm compared to ℓ^2 norm space for solving the quadratic minimization problem.

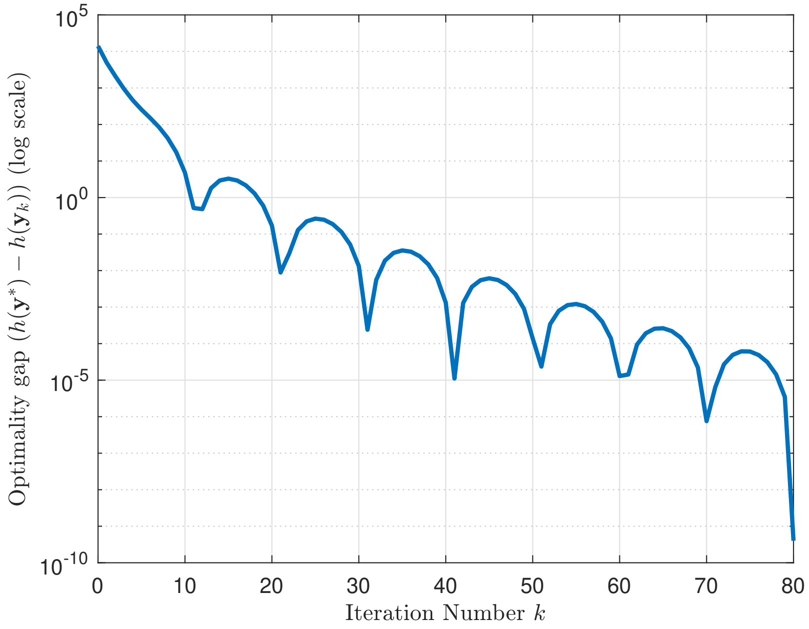

Figure 1: Variation of optimality gap with iteration in **Case 1**

6 Proposed algorithm

6.1 Generalized Conceptual Dual Implicit Accelerated Gradient Descent

We present the extension of the conceptual version of Dual Implicit Accelerated Gradient algorithm (C-DIAG) [9] for the general norm in Algorithm 2

Algorithm 2 Generalized Conceptual Dual Implicit Accelerated Gradient Descent (GC-DIAG) for strongly-convex-concave programming

```
   Input: $g$, $D_\psi$, $\mu_\psi$, $L$, $\sigma$, $\mathbf{x}_0$, $\mathbf{y}_0$, $K$,
   Output: $\bar{\mathbf{x}}_K$ $\mathbf{y}_K$
1: Set $\beta > L$, $\mathbf{z}_0 \leftarrow \mathbf{y}_0$
2: for $k = 0, 1, ..., K$ do
3:     $\tau_k \leftarrow \frac{2}{k+2}$, $\eta_k \leftarrow \frac{k+1}{2\beta}\mu_\psi$, $\mathbf{w}_k \leftarrow (1-\tau_k)\,\mathbf{y}_k + \tau_k \mathbf{z}_k$
4:     Choose $\mathbf{x}_{k+1}, \mathbf{y}_{k+1}$, ensuring:
           $$g(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \min_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_{k+1}),\; \mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y}\in\mathcal{Y}} \left\{ -\langle \nabla_{\mathbf{y}} g(\mathbf{x}_{k+1}, \mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \rangle + \frac{\beta}{2}\left\|\mathbf{y} - \mathbf{w}_k\right\|^2 \right\}$$
5:     $\mathbf{z}_{k+1} \leftarrow \arg\min_{\mathbf{z}\in\mathcal{Y}} \{-\eta_k \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{k+1}, \mathbf{w}_k), \mathbf{z}\rangle + D_\psi(\mathbf{z}\|\mathbf{z}_k)\}$,    $\bar{\mathbf{x}}_{k+1} \leftarrow \frac{2}{(k+1)(k+2)} \sum_{i=1}^{k+1} i\mathbf{x}_i$
6: end for
7: return $\bar{\mathbf{x}}_K, \mathbf{y}_K$
```

Convergence Analysis

The updates $\mathbf{y}_{k+1}$ in Step 4 and $\mathbf{z}_{k+1}$ in Step 5 correspond to the Accelerated Gradient Ascent update of $g(\mathbf{x}_{k+1}, .)$. Hence, we use our analysis for Algorithm 1. From Lemma 2, we conclude that $\forall \mathbf{y} \in \mathcal{Y}$,

$$\underbrace{(k+1)\,(k+2)\,(g(\mathbf{x}_{k+1}, \mathbf{y}) - g(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_{k+1})}_{\Phi_{k+1}}$$

$$
\begin{aligned}
&\overset{(1)}{\leq} (k)\,(k+1)\,(g(\mathbf{x}_{k+1},\mathbf{y}) - g(\mathbf{x}_{k+1},\mathbf{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_k) \\
&\overset{(2)}{=} (k)\,(k+1)\,(g(\mathbf{x}_{k+1},\mathbf{y}) - g(\mathbf{x}_k,\mathbf{y})) + (k)\,(k+1)\,(g(\mathbf{x}_k,\mathbf{y}) - g(\mathbf{x}_{k+1},\mathbf{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_k) \\
&\overset{(3)}{\leq} (k)\,(k+1)\,(g(\mathbf{x}_{k+1},\mathbf{y}) - g(\mathbf{x}_k,\mathbf{y})) + \underbrace{(k)\,(k+1)\,(g(\mathbf{x}_k,\mathbf{y}) - g(\mathbf{x}_k,\mathbf{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_k)}_{\Phi_k} \\
&\overset{(4)}{\leq} (k)\,(k+1)\,g(\mathbf{x}_{k+1},\mathbf{y}) - \sum_{i=1}^{k}(2i)(g(\mathbf{x}_i,\mathbf{y})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_0),
\end{aligned} \tag{33}
$$

where inequality (1) follows from Lemma 2. Equality (2) follows by adding and subtracting $k(k+1)g(\mathbf{x}_k,\mathbf{y})$. Inequality (3) follows from the update rule $g(\mathbf{x}_k,\mathbf{y}_k) = \min_{\mathbf{x}} g(\mathbf{x},\mathbf{y}_k)$ in Step 4 of Algorithm 2 and hence $g(\mathbf{x}_k,\mathbf{y}_k) \leq g(\mathbf{x}_{k+1},\mathbf{y}_k)$. Inequality (4) follows from the recurrence relation established in inequality (3). Now, we write

$$
(k+1)\,(k+2)\,(g(\mathbf{x}_{k+1},\mathbf{y}) - g(\mathbf{x}_{k+1},\mathbf{y}_{k+1})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_{k+1}) \tag{34}
$$

$$
\leq (k)\,(k+1)\,g(\mathbf{x}_{k+1},\mathbf{y}) - \sum_{i=1}^{k}(2i)(g(\mathbf{x}_i,\mathbf{y})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_0). \tag{35}
$$

Rearranging the terms, we get

$$
\sum_{i=1}^{k+1}(2i)(g(\mathbf{x}_i,\mathbf{y})) - (k+1)(k+2)g(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \leq \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_0) - \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_{k+1}) \tag{36}
$$

$$
\sum_{i=1}^{k+1}(2i)(g(\mathbf{x}_i,\mathbf{y})) - (k+1)(k+2)g(\mathbf{x},\mathbf{y}_{k+1}) \overset{(5)}{\leq} \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_0) \tag{37}
$$

$$
g(\bar{\mathbf{x}}_{k+1},\mathbf{y})) - g(\mathbf{x},\mathbf{y}_{k+1}) \overset{(6)}{\leq} \frac{4\beta}{\mu_\psi(k+1)(k+2)} D_\psi(\mathbf{y}\|\mathbf{z}_0) \tag{38}
$$

$$
\max_{\mathbf{y}\in\mathcal{Y}} g(\bar{\mathbf{x}}_{k+1},\mathbf{y})) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_{k+1}) \overset{(7)}{\leq} \frac{4\beta}{\mu_\psi(k+1)(k+2)} D_\psi(\mathbf{y}\|\mathbf{z}_0) \tag{39}
$$

where inequality (5) follows by dropping the term $\frac{4\beta}{\mu_\psi} D_\psi(\mathbf{y}\|\mathbf{z}_{k+1})$ since Bregman divergence is positive and from the fact that $g(\mathbf{x},\mathbf{y}_{k+1}) \geq g(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \forall \mathbf{x} \in \mathcal{X}$. Inequality (6) follows by defining a convex combination of $\bar{\mathbf{x}}_{k+1} = \frac{1}{(k+1)(k+2)} \sum_{i=1}^{k+1}(2i)\mathbf{x}_i$ and from the fact that $g(.,\mathbf{y})$ is convex for every $\mathbf{y}$. Since inequality (6) is satisfied for arbitrary $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, we maximise $\mathbf{y}$ over $\mathcal{Y}$ and minimize $\mathbf{x}$ over $\mathcal{X}$ to get inequality (7). From inequality (7) and Definition 3, we conclude that algorithm 2 gives a $\mathcal{O}(1/k^2)$ convergence rate for the primal dual gap.

6.2 Generalized Dual Implicit Accelerated Gradient Descent

The proof of convergence rate of outer loop of G-DIAG follows along the lines of the potential-function based proof of Nesterov's AGD with general norm (Section 5.5). We would construct an appropriate potential-function and then show that its value decreases monotonically over the iterations. By assuming that $\texttt{Imp} - \texttt{STEP}$ is guaranteed to converge, we can re-write each iteration of the G-DIAG algorithm as

$$
\tau_k = ?, \quad \eta_k = ? \tag{41}
$$

$$
\mathbf{w}_k = (1-\tau_k)\,\mathbf{y}_k + \tau_k \mathbf{z}_k \tag{42}
$$

$$
\mathbf{y}_{k+1} = \arg\min_{\mathbf{y}\in\mathcal{Y}} \left\{ -\langle \nabla_{\mathbf{y}} h_{k+1}(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{w}_k\|^2 \right\} \tag{43}
$$

$$
\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}\in\mathcal{Y}} \left\{ -\eta_k \langle \nabla_{\mathbf{y}} h_{k+1}, \mathbf{z} \rangle + D_\psi(\mathbf{z}\|\mathbf{z}_k) \right\} \tag{44}
$$

where $h_{k+1}(\mathbf{y}) := g(\mathbf{x}_{k+1},\mathbf{y})$ such that $g(\mathbf{x}_{k+1},\mathbf{y}_{k+1}) \leq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x},\mathbf{y}_{k+1}) + \epsilon_{\text{step}}^{(k+1)}$. Clearly, G-DIAG executes the k-th step iteration of the accelerated gradient descent with general norm (Algortihm 1) for the concave function

Algorithm 3 Generalized Dual Implicit Accelerated Gradient Descent (G-DIAG) for strongly-convex-concave programming

```
Input: $g$, $D_\psi$, $\mu_\psi$, $L$, $\sigma$, $\mathbf{x}_0$, $\mathbf{y}_0$, $K$, $\left\{\epsilon_{\text{step}}^{(k)}\right\}_{k=1}^{K}$
Output: $\bar{\mathbf{x}}_K$ $\mathbf{y}_K$
Set $\beta \leftarrow L$, $\mathbf{z}_0 \leftarrow \mathbf{y}_0$
for $k = 0, 1, ..., K$ do
    $\tau_k \leftarrow \frac{2}{k+2}$, $\eta_k \leftarrow \frac{k+1}{2\beta}\psi$, $\mathbf{w}_k \leftarrow (1-\tau_k)\,\mathbf{y}_k + \tau_k \mathbf{z}_k$
    $\mathbf{x}_{k+1}, \mathbf{y}_{k+1} \leftarrow \mathtt{Imp}-\mathtt{STEP}(g, L, \sigma, \mathbf{x}_0, \mathbf{w}_k, \beta, \epsilon_{\text{step}}^{(k+1)})$, ensuring:
    $g(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \leq \min_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_{k+1}) + \epsilon_{\text{step}}^{(k+1)}$, $\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y}\in\mathcal{Y}} \left\{ -\left\langle \nabla_{\mathbf{y}} g(\mathbf{x}_{k+1}, \mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \right\rangle + \frac{\beta}{2}\left\|\mathbf{y} - \mathbf{w}_k\right\|^2 \right\}$
    $\mathbf{z}_{k+1} \leftarrow \arg\min_{\mathbf{z}\in\mathcal{Y}} \{-\eta_k \left\langle \nabla_{\mathbf{y}} g(\mathbf{x}_{k+1}, \mathbf{w}_k), \mathbf{z} \right\rangle + D_\psi(\mathbf{z}\|\mathbf{z}_k)\}$,    $\bar{\mathbf{x}}_{k+1} \leftarrow \frac{2}{(k+1)(k+2)} \sum_{i=1}^{k+1} i\mathbf{x}_i$
end for
return $\bar{\mathbf{x}}_K, \mathbf{y}_K$

$\mathtt{Imp}-\mathtt{STEP}(g, L, \sigma, \mathbf{x}_0, \mathbf{w}, \beta, \epsilon_{\text{step}}^{(k+1)})$:
    Set $R \leftarrow ?$, $\epsilon_{\text{agd}} \leftarrow ?$, $\mathbf{y}_0 \leftarrow \mathbf{w}$
    for $r = 0, 1, ..., R$ do
        Starting at $\mathbf{x}_0$ use generalized AGD (Algorithm 1 with $-g(., \mathbf{y}_r)$ to compute $\mathbf{x}_r$ such that:
            $$g(\hat{\mathbf{x}}_r, \mathbf{y}_r) \leq \min_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}_r) + \epsilon_{\text{agd}}, \tag{40}$$
        $\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y}\in\mathcal{Y}} \left\{ -\left\langle \nabla_{\mathbf{y}} g(\hat{\mathbf{x}}_r, \mathbf{w}), \mathbf{y} - \mathbf{w} \right\rangle + \frac{\beta}{2}\left\|\mathbf{y} - \mathbf{w}\right\|^2 \right\}$
    end for
    return $\hat{\mathbf{x}}_R, \mathbf{y}_{R+1}$
```

$h_{k+1} = g(\mathbf{x}_{k+1}, .)$. As in Equation (16), we can define the potential function for iteration j using the concave function h_k and an arbitrary reference point $\tilde{\mathbf{y}} \in \mathcal{Y}$ as

$$\Phi^{h_k}(j) = j(j+1)\left(h_k(\tilde{\mathbf{y}}) - h_k(\mathbf{y}_j)\right) + \frac{4\beta}{\mu_\psi} D(\tilde{\mathbf{y}}||\mathbf{z}_j) \tag{45}$$

From Lemma 2, using $\beta > L$, the potential function $\Phi^{h_{k+1}}(k)$ decreases at the step k of the algorithm. Hence,

$$\begin{aligned}
\Phi^{h_{k+1}}(k+1) \leq \Phi^{h_{k+1}}(k)& \\
&= k(k+1)\left(h_{k+1}(\tilde{\mathbf{y}}) - h_{k+1}(\mathbf{y}_k)\right) + \frac{4\beta}{\mu_\psi} D(\tilde{\mathbf{y}}||\mathbf{z}_k) \\
&= k(k+1)\left(h_k(\tilde{\mathbf{y}}) - h_k(\mathbf{y}_k)\right) + \frac{4\beta}{\mu_\psi} D(\tilde{\mathbf{y}}||\mathbf{z}_k) + \\
&\quad k(k+1)\left(h_{k+1}(\tilde{\mathbf{y}}) - h_k(\tilde{\mathbf{y}})\right) + k(k+1)\left(h_k(\mathbf{y}_k) - h_{k+1}(\mathbf{y}_k)\right) \\
&= \Phi^{h_k}(k) + k(k+1)\left(g(\mathbf{x}_{k+1}, \tilde{\mathbf{y}}) - g(\mathbf{x}_k, \tilde{\mathbf{y}})\right) + k(k+1)\left(g(\mathbf{x}_k, \mathbf{y}_k) - g(\mathbf{x}_{k+1}, \mathbf{y}_k)\right) \\
&\overset{(a)}{\leq} \Phi^{h_k}(k) + k(k+1)\left(g(\mathbf{x}_{k+1}, \tilde{\mathbf{y}}) - g(\mathbf{x}_k, \tilde{\mathbf{y}})\right) + k(k+1)\epsilon_{\text{step}}^{(k)}
\end{aligned}$$

where (a) follows provided Imp $-$ STEP converges and $g(\mathbf{x}_k, \mathbf{y}_k) - g(\mathbf{x}_{k+1}, \mathbf{y}_k) \leq g(\mathbf{x}_k, \mathbf{y}_k) - \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_k) \leq \epsilon_{\text{step}}^{(k)}$. Carrying out summation for $k = 0, \ldots, K-1$ and rearranging the terms gives

$$\begin{aligned}
\Phi^{h_0}(0) + \sum_{k=1}^{K-1} \epsilon_{\text{step}}^{(k)} &\geq \sum_{k=1}^{K-1} 2kg(\mathbf{x}_k, \tilde{\mathbf{y}}) + \Phi^{h_K}(K) - (K-1)Kg(\mathbf{x}_K, \tilde{\mathbf{y}}) \\
&\geq \sum_{k=1}^{K-1} g(\mathbf{x}_k, \tilde{\mathbf{y}}) + K(K+1)\left(g(\mathbf{x}_K, \tilde{\mathbf{y}}) - g(\mathbf{x}_K, \mathbf{y}_K)\right) + \frac{4\beta}{\mu_\psi} D(\tilde{\mathbf{y}}||\mathbf{z}_K) - (K-1)Kg(\mathbf{x}_K, \tilde{\mathbf{y}})
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \sum_{k=1}^{K} g(\mathbf{x}_k, \tilde{\mathbf{y}}) - K(K+1) g(\mathbf{x}_K, \mathbf{y}_K) \\
&\overset{(b)}{\geq} K(K+1)\left[g(\bar{\mathbf{x}}_K, \tilde{\mathbf{y}}) - g(\mathbf{x}_K, \mathbf{y}_K)\right] \\
&\overset{(c)}{\geq} K(K+1)\left[g(\bar{\mathbf{x}}_k, \tilde{\mathbf{y}}) - g(\tilde{\mathbf{x}}, \mathbf{y}_K) - \epsilon_{\text{step}}^{(K)}\right] \qquad (46)
\end{aligned}
$$

where (a) holds because $D(\tilde{\mathbf{y}}||\mathbf{z}_K) \geq 0$, (b) uses the fact that $\bar{\mathbf{x}}_K = \frac{1}{K+1}\sum_{k=1}^{K} 2i\mathbf{x}_i$ and (c) follows provided $\text{Imp} - \text{STEP}$ converges and $g(\mathbf{x}_K, \mathbf{y}_K) \leq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_K) + \epsilon_{\text{step}}^{(K)} \leq g(\tilde{\mathbf{x}}, \mathbf{y}_K) + \epsilon_{\text{step}}^{(K)}$. Rearranging the terms, we get

$$
\begin{aligned}
g(\bar{\mathbf{x}}_k, \tilde{\mathbf{y}}) - g(\tilde{\mathbf{x}}, \mathbf{y}_K) &\leq \frac{\Phi^{h_0}(0)}{K(K+1)} + \sum_{k=1}^{K} \frac{k(k+1)}{K(K+1)} \epsilon_{\text{step}}^{(k)} \\
&= \frac{4\beta D_\psi(\tilde{\mathbf{y}}||\mathbf{y}_0)}{\psi K(K+1)} + \sum_{k=1}^{K} \frac{k(k+1)}{K(K+1)} \epsilon_{\text{step}}^{(k)} \qquad (47)
\end{aligned}
$$

The final result follows by taking maximum over $\tilde{\mathbf{x}}$ and minimum over $\tilde{\mathbf{y}}$ repectively. If we set $\epsilon_{\text{step}} = \frac{L\Omega}{\psi k^3(k+1)}$ and ignore the effect of $\epsilon_{\text{step}}^{(k)}$ on the oracle complexity of $\texttt{Imp} - \texttt{STEP}$, we get

$$
\begin{aligned}
\max_{\tilde{\mathbf{y}}\in\mathcal{Y}} g(\bar{\mathbf{x}}_k, \tilde{\mathbf{y}}) - \min_{\tilde{\mathbf{x}}\in\mathcal{X}} g(\tilde{\mathbf{x}}, \mathbf{y}_K) &\leq \frac{4L\Omega}{\psi K(K+1)} + \frac{L\Omega}{K(K+1)} \sum_{k=1}^{K} \frac{1}{k^2} \\
&\overset{(a)}{\leq} \frac{6L\Omega}{\psi K(K+1)} \qquad (48)
\end{aligned}
$$

where (a) follows because $\sum_{k=1}^{K} \frac{1}{k^2} < 2 - \frac{1}{K}$. Note that this is not the overall oracle complexity of the G-DIAG algorithm cause as it ignores the number of oracle calls in the $\texttt{Imp} - \texttt{STEP}$.

7 Problems faced with the proof

We faced the following problems during our attempt to construct the proof of G-DIAG

1. In the $\texttt{Imp} - \texttt{Step}$ of the original DIAG algorithm [9], AGD (with ℓ^2 norm) is used to efficiently calculate (in logarithmic number of steps) an estimate for $\mathbf{x}_r = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_r) + \epsilon_{\text{agd}}$. The proof uses the guarantee on AGD for strongly convex case in [2, Equation (5.68)]. But in our setting, strong convexity is defined with respect to some arbitrary norm. We have not found any literature that tackles this problem using AGD. It may happen that, for arbitrary norm, finding an ϵ_{agd} estimate $\mathbf{x}_r$ would not be possible in linear time. As an example, we found a blog [10] that discusses the case of mirror descent for the strongly convex case with respect to a general norm. They note that the improved oracle complexity (due to strong convexity) is $\mathcal{O}(1/k)$ only. So, it may not be possible to achieve the linear convergence in an arbitrary norm case.
2. In order to show that the $\texttt{Imp} - \texttt{Step}$ converges, we need to show the updates satisfy some fixed point equation. Let

$$
\mathbf{x}^*(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}) \qquad (49)
$$

$$
\mathbf{y}^+ = \arg\min_{\mathbf{y}'\in\mathcal{Y}} \left\{ - \langle \nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}), \mathbf{w}), \mathbf{y}' - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{y}' - \mathbf{w}\|^2 \right\} \qquad (50)
$$

Then for proving that there exists a fixed point of the iterations of the $\texttt{Imp} - \texttt{Step}$, we attempt to show that the operation $\mathbf{y}^+$ is a contraction,

$$
\left\|\mathbf{y}_1^+ - \mathbf{y}_2^+\right\| \leq \alpha \left\|\mathbf{y}_1 - \mathbf{y}_2\right\| \qquad (51)
$$

for some $\alpha < 1$. Note that $\left\|\mathbf{y}_1^+ - \mathbf{y}_2^+\right\|$ equals

$$\alpha \left\| \arg\min_{\mathbf{y}' \in \mathcal{Y}} \left\{ - \langle \nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}_1), \mathbf{w}), \mathbf{y}' - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{y}' - \mathbf{w}\|^2 \right\} - \arg\min_{\mathbf{y}' \in \mathcal{Y}} \left\{ - \langle \nabla_{\mathbf{y}} g(\mathbf{x}^*(\mathbf{y}_2), \mathbf{w}), \mathbf{y}' - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{y}' - \mathbf{w}\|^2 \right\} \right\|$$

In [9], due to the ℓ_2 norm space, they proceed using Pythogoras theorem of the projection operator. But, that is not possible in our case. One possible solution to this is to use Bregman projections instead and relative smoothness and convexity definitions. But we could not prove AGD convergence using relative smoothness and convexity as that would be required for the proof. We discuss this further in the appendix.

7.1 Attempted solution to problem 1

In case of mirror descent, the restarting strategy discussed in [10] improves the convergence rate from $\mathcal{O}(1/\sqrt{k})$ to $\mathcal{O}(1/k)$ through introduction of strong convexity (or concavity in our case). We try to come up with the convergence rate for generalized AGD with strong concavity following a similar procedure.

Using Lemma 2, the oracle complexity for generalized AGD can be found by carrying out a telescoping sum. We state the result in the following theorem

Theorem 1. *Suppose $h(.)$ is a L-smooth function and the parameters of Algorithm 1 are chosen as per Lemma 2, then the following holds*

$$h(\mathbf{y}^*) - h(\mathbf{y}_T) \leq \frac{4\beta}{\mu_\psi} \cdot \frac{D_\psi(\mathbf{y}^* || \mathbf{y}_0)}{T(T+1)} \tag{52}$$

If we also assume $\Omega = \max_{\mathbf{y} \in \mathcal{Y}} D_\psi(\mathbf{y}||\mathbf{y}_0)$, then the following bound would hold

$$h(\mathbf{y}^*) - h(\mathbf{y}_T) \leq \frac{4\beta}{\mu_\psi} \cdot \frac{\Omega}{T(T+1)} \tag{53}$$

Now, assume that $h(.)$ is also σ-strongly concave with respect to some norm $\|.\|$. Further, $\psi(.)$ is chosen such that it is μ_ψ-strongly convex on the whole $\mathbb{R}^d$, *instead of only on $\mathcal{Y}$*. For any $R > 0$ and $\mathbf{u}$, define $\psi_{R,\mathbf{u}}(\mathbf{y}) := \psi(R^{-1}(\mathbf{y} - \mathbf{u}))$. Let $D_{\psi,R,\mathbf{z}}(.||.)$ denote the corresponding Bregman divergence. The following corollary is crucial to get a bound for generalized AGD with strong convexity.

Corollary 1.1. *Suppose*

$$\Omega = \max\left\{D_\psi(\mathbf{v}||\mathbf{0}) | \, \|\mathbf{v}\| \leq 1, \mathbf{v} \in \mathbb{R}^d\right\}, \quad R_0 = \|\mathbf{y}^* - \mathbf{y}_0\|$$

If we apply Algorithm 1 with $\eta_k = \frac{(k+1)}{2R_0\beta}$, $\tau_k = \frac{2}{k+2}$, learning rate $\frac{1}{R_0\beta}$ for some $\beta > L$ and Bregman divergence $D_{\psi,R,\mathbf{y}_0}(.||.)$ for T iterations. Then, the following bounds hold

$$h(\mathbf{y}^*) - h(\mathbf{y}_T(R_0, \mathbf{y}_0)) \leq \frac{4R_0\beta}{\mu_\psi} \cdot \frac{\Omega}{T(T+1)} \tag{54}$$

$$\|\mathbf{y}^* - \mathbf{y}_T(R_0, \mathbf{y}_0)\|^2 \leq \frac{8R_0\beta}{\mu_\psi \mu} \cdot \frac{\Omega}{T(T+1)} \tag{55}$$

where $\mathbf{y}^*$ is the unique maximizer of $h(.)$ on $\mathcal{Y}$.

Proof. Consider the norm $\|.\|_{R_0} := R_0^{-1} \|.\|$. Note that $h(.)$ is R_0L-smooth and $\psi_{R,\mathbf{u}}(.)$ is μ_ψ-strongly convex with respect to $\|.\|_{R_0}$. In this case, Ω would become

$$D_{\psi,R_0,\mathbf{y}_0}(\mathbf{y}^*||\mathbf{y}_0) = D_\psi(R_0^{-1}(\mathbf{y}^* - \mathbf{y}_0)||\mathbf{0}) = \max\left\{D_\psi(\mathbf{v}||\mathbf{0}) | \, \|\mathbf{v}\| \leq 1, \mathbf{v} \in \mathbb{R}^d\right\} := \Omega$$

The bound in (54) follows directly from Theorem 1. The bound in Equation (55) is obtained from (54) using strong-concavity of $h(.)$ and the optimality of $\mathbf{y}^*$.

$$h(\mathbf{y}^*) - h(\mathbf{y}_T) \geq \langle \nabla h(\mathbf{y}^*), \mathbf{y}^* - \mathbf{y}_T \rangle + \frac{\mu}{2} \|\mathbf{y}^* - \mathbf{y}_T\|^2, \quad \langle \nabla h(\mathbf{y}^*), \mathbf{y}^* - \mathbf{y}_T \rangle \geq 0$$

□

The error bounds given by Equations (54) (55) depend on R_0 with smaller R_0 giving smaller error bounds. Also, the bound of the distance between $\mathbf{y}^*$ and $\mathbf{y}_T$ (55) is strictly decreasing with iterations T. These observations can be used to design the following restarting strategy.

1. Set $\mathbf{y}_0 \in \mathcal{Y}$, $l = 0$
2. Set T_l such that $\|\mathbf{y}^* - \mathbf{y}_{T_l}(R_l, \mathbf{y}_l)\|^2 \leq 2^{-1} R_l^2$.
3. Compute $\mathbf{y}_{l+1} = \mathbf{y}_{T_l}(R_l, \mathbf{y}_l)$ using generalized AGD as per Corollary 1.1.
4. Set $R_{l+1}^2 = 2^{-1} R_l^2$, $l = l + 1$. Go to step 2.

By Corollary 1.1, it suffices to choose T_l such that

$$\frac{8R_l\beta}{\mu_\psi \mu} \cdot \frac{\Omega}{T_l(T_l+1)} \leq \frac{8R_l\beta}{\mu_\psi \mu} \cdot \frac{\Omega}{T_l^2} \leq 2^{-1} R_l^2$$

$$\implies T_l = \left\lceil \sqrt{\frac{16\beta\Omega}{R_l \mu_\psi \mu}} \right\rceil \tag{56}$$

The total number of AGD iterations required to get $\mathbf{y}_L$ is defined as $M_L = \sum_{l=0}^{L-1} T_l$.

Proposition 1. *Let L^* be the largest L such that*

$$L \geq \sqrt{\frac{8\beta\Omega}{R_0 \mu_\psi \mu}} 2^{(L+1)/4}$$

Then, the proposed restarting strategy guarantees the following bound

$$h(\mathbf{y}^*) - h(\mathbf{y}_L) \leq 2^{-(0.5M_L+1)} \mu R_0^2, \quad \textit{for } L \leq L^* \tag{57}$$

$$\|\mathbf{y}^* - \mathbf{y}_L\|^2 \leq 2^{-0.5M_L} R_0^2, \quad \textit{for } L \leq L^* \tag{58}$$

and

$$h(\mathbf{y}^*) - h(\mathbf{y}_L) \leq \frac{1024\beta^2\Omega^2}{\mu_\psi^2 \mu M_L^4}, \quad \textit{for } L > L^* \tag{59}$$

$$\|\mathbf{y}^* - \mathbf{y}_L\|^2 \leq \frac{2048\beta^2\Omega^2}{\mu_\psi^2 \mu^2 M_L^4}, \quad \textit{for } L > L^* \tag{60}$$

Proof. Using the proposed restarting strategy, it follows from Corollary 1.1 that

$$h(\mathbf{y}^*) - h(\mathbf{y}_L) \leq 2^{-(L+1)} \mu R_0^2 \tag{61}$$

$$\|\mathbf{y}^* - \mathbf{y}_L\|^2 \leq 2^{-L} R_0^2 \tag{62}$$

By the choice of T_l given in Equation (56), it holds that

$$M_L \leq L + \sum_{l=0}^{L-1} \sqrt{\frac{16\beta\Omega}{R_l \mu_\psi \mu}} = L + \sum_{l=0}^{L-1} \sqrt{\frac{16\beta\Omega}{R_0 \mu_\psi \mu}} 2^{l/4} \leq L + \sqrt{\frac{8\beta\Omega}{R_0 \mu_\psi \mu}} 2^{(L+1)/4}$$

Therefore, depending on value of L, the following holds

$$M_L \leq 2L, \quad \text{for } L \leq L^* \tag{63}$$

$$M_L \leq \sqrt{\frac{32\beta\Omega}{R_0 \mu_\psi \mu}} 2^{(L+1)/4}, \quad \text{for } L > L^* \tag{64}$$

The bounds given in Equations (57)-(60) follow by eliminating L from Equations (61), (62) using the relations in Equations (63), (64). □

We observe that the introduction of strong convexity improves the convergence rate of generalized AGD from $\mathcal{O}(1/k^2)$ to $\mathcal{O}(1/k^4)$. Note that as expected, this is significantly better than the $\mathcal{O}(1/k)$ convergence rate obtained by applying mirror descent.

8 Conclusion

In this term project, we extended the framework of Conceptual Dual Implicit Accelerated Gradient Descent to arbitrary norm case and proved $\mathcal{O}(1/k^2)$ convergence rate using Bregman divergence framework. A key intermediate step involved proving convergence guarantee for Nesterov's AGD for a strongly convex and smooth function with respect to an arbitrary norm. We improved the convergence to $\mathcal{O}(\frac{1}{k^4})$ using the restarting strategy. We plan to use the notion of relative smoothness and strong convexity to prove the contraction bound required for the inexact version of Dual Implicit Accelerated Gradient Descent.

A Attempt to analyze Nesterov's AGD using notion of relative smoothness

We first state the definition of relative smoothness as discussed in the course.

Definition 4. *The function $f(.)$ is $L-$ smooth relative to the reference function $\psi(.)$ if for any x and y, there is a scalar L for which*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_\psi(y \| x) \tag{65}$$

Algorithm 4 Nesterov's accelerated gradient ascent for Non-Euclidean space (Concept of relative smoothness)

```
    Input: Smooth concave function $h(.)$, learning rate $\frac{1}{\beta}$, Bregman divergence $D_\psi(.\|.)$, initial point $\mathbf{y}_0$ and $\mathbf{z}_0$
    Output: $\mathbf{y}_K$
for $k = 0, 1, ..., K$ do
    $\mathbf{w}_k \leftarrow (1 - \tau_k)\mathbf{y}_k + \tau_k \mathbf{z}_k$    (66)
    $\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y} \in \mathcal{Y}} \{-\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k \rangle + \beta D_\psi(\mathbf{y} \| \mathbf{w}_k)\}$    (67)
    $\mathbf{z}_{k+1} \leftarrow \arg\min_{\mathbf{z} \in \mathcal{Y}} \{-\eta_k \langle \nabla h(\mathbf{w}_k), \mathbf{z} \rangle + D_\psi(\mathbf{z} \| \mathbf{z}_k)\}$    (68)
end for
```

We consider a function $h(.)$ which is L-smooth relative to $\psi(.)$. We modify the algorithm 1 by replacing the $\frac{\beta}{2}\left\|\mathbf{y} - \mathbf{w}_k\right\|^2$ term with $\beta D_\psi(\mathbf{y} \| \mathbf{w}_k)$ shown in algorithm 4. For the same choice of $\beta > L$, $\eta_k = \frac{(k+1)}{2\beta}\mu_\psi$ and $\tau_k = \frac{2}{k+2}$, we attempt to prove $\Phi(k+1) \leq \Phi(k)$ where $\Phi(k)$ is defined in (16).

Using (16), the potential difference can be written as

$$\begin{aligned}\Phi(k+1) - \Phi(k) = (k+1)(k+2)&\underbrace{(h(\mathbf{w}_k) - h(\mathbf{y}_{k+1}))}_{(a)} \\ \underbrace{-k(k+1)(h(\mathbf{w}_k) - h(\mathbf{y}_k)) + 2(k+1)(h(\mathbf{y}) - h(\mathbf{w}_k))}_{(b)} &+ \frac{4\beta}{\mu_\psi}\underbrace{(D_\psi(\mathbf{y} \| \mathbf{z}_{k+1}) - D_\psi(\mathbf{y} \| \mathbf{z}_k))}_{(c)}\end{aligned} \tag{69}$$

The term (c) can be bounded as

$$\begin{aligned}D_\psi(\mathbf{y} \| \mathbf{z}_{k+1}) - D_\psi(\mathbf{y} \| \mathbf{z}_k) &= (\psi(\mathbf{y}) - \psi(\mathbf{z}_{k+1}) - \langle \nabla\psi(\mathbf{z}_{k+1}), \mathbf{y} - \mathbf{z}_{k+1} \rangle) - (\psi(\mathbf{y}) - \psi(\mathbf{z}_k) - \langle \nabla\psi(\mathbf{z}_k), \mathbf{y} - \mathbf{z}_k \rangle) \\ &= \psi(\mathbf{z}_k) - \psi(\mathbf{z}_{k+1}) + \langle \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{z}_k \rangle + \langle \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{y} \rangle \\ &= -D_\psi(\mathbf{z}_{k+1} \| \mathbf{z}_k) + \underbrace{\langle \nabla\psi(\mathbf{z}_{k+1}) - \nabla\psi(\mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{y} \rangle}_{(d)} \\ &\leq -D_\psi(\mathbf{z}_{k+1} \| \mathbf{z}_k) + \langle \eta_k \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y} \rangle\end{aligned} \tag{70}$$

where the last inequality is similar to that proved in Lemma 2. Note the difference between (18) and (70).

The term (a) can be bounded as

$$h(\mathbf{y}_{k+1}) - h(\mathbf{w}_k) \overset{(1)}{\geq} \langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k \rangle - L D_\psi(\mathbf{y}_{k+1} \| \mathbf{w}_k) \overset{(2)}{\geq} \langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k \rangle - \beta D_\psi(\mathbf{y}_{k+1} \| \mathbf{w}_k) \tag{71}$$

Here, inequality (1) follows from the fact that $(-h(\mathbf{x}))$ is L-smooth relative to $\psi(.)$, inequality (2) follows from the choice of $\beta > L$. Note the difference between (22) and (71).

From the update in (67) in Algorithm 4, $\mathbf{y}_{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \{\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{w}_k\rangle - \beta D_\psi(\mathbf{y}\|\mathbf{w}_k)\}$. We know that $\mathbf{y}_k \in \mathcal{Y}$ and $\mathbf{z}_{k+1} \in \mathcal{Y}$. So, a convex combination $\mathbf{v} = (1-\tau_k)\mathbf{y}_k + \tau_k \mathbf{z}_{k+1} \in \mathcal{Y}$. Hence, we can write

$$\begin{aligned}\langle \nabla h(\mathbf{w}_k), \mathbf{y}_{k+1} - \mathbf{w}_k\rangle - \beta D_\psi(\mathbf{y}_{k+1}\|\mathbf{w}_k) &\geq \langle \nabla h(\mathbf{w}_k), \mathbf{v} - \mathbf{w}_k\rangle - \beta D_\psi(\mathbf{v}\|\mathbf{w}_k) \\ &\overset{(3)}{=} \tau_k \langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{z}_k\rangle - \beta D_\psi(\mathbf{v}\|\mathbf{w}_k) \qquad (72)\end{aligned}$$

where equality (3) follows from the update rule (66) in Algorithm 4.

The term (b) can be bounded similar to Lemma 2 as

$$-k(k+1)(h(\mathbf{w}_k) - h(\mathbf{y}_k)) + 2(k+1)(h(\mathbf{y}) - h(\mathbf{w}_k)) \leq 2(k+1)\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{z}_k\rangle, \qquad (73)$$

We now substitute bounds (70), (72) and (73) in (69) to get

$$\begin{aligned}\Phi(k+1) - \Phi(k) &\leq (k+1)(k+2)\left(-\tau_k \langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{z}_k\rangle + \beta D_\psi(\mathbf{v}\|\mathbf{w}_k)\right) \\ &\quad + 2(k+1)\langle \nabla h(\mathbf{w}_k), \mathbf{y} - \mathbf{z}_k\rangle + \frac{4\beta}{\mu_\psi}\left(-D_\psi(\mathbf{z}_{k+1}\|\mathbf{z}_k) + \langle \eta_k \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle\right) \\ &\overset{(???)}{\leq} \underbrace{(k+1)(k+2)\beta D_\psi(\mathbf{v}\|\mathbf{w}_k) - \frac{4\beta}{\mu_\psi} D_\psi(\mathbf{z}_{k+1}\|\mathbf{z}_k)}_{\text{How to prove } \leq 0} + \underbrace{\left(-2(k+1) + \frac{4\beta}{\mu_\psi}\eta_k\right)\langle \nabla h(\mathbf{w}_k), \mathbf{z}_{k+1} - \mathbf{y}\rangle}_{=0 \text{ by the choice of } \eta_k = \frac{(k+1)}{2\beta}\mu_\psi} \\ &\qquad (74)\end{aligned}$$

References

[1] Kwangjun Ahn and Suvrit Sra. "From Nesterov's Estimate Sequence to Riemannian Acceleration". In: *arXiv preprint arXiv:2001.08876* (2020).

[2] Nikhil Bansal and Anupam Gupta. "Potential-function proofs for first-order methods". In: *arXiv preprint arXiv:1712.04581* (2017).

[3] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[4] Yuxin Chen. *Y. Chen, "Notes on large scale optimization for data science,"*. URL: `https://www.princeton.edu/~yc5/ele522_optimization/lectures.html`.

[5] Haihao Lu, Robert M Freund, and Yurii Nesterov. "Relatively smooth convex optimization by first-order methods, and applications". In: *SIAM Journal on Optimization* 28.1 (2018), pp. 333–354.

[6] Arkadi Nemirovski. "Prox-Method with Rate of Convergence O (1/ t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems". In: *SIAM Journal on Optimization* 15 (Jan. 2004), pp. 229–251. DOI: `10.1137/S1052623403425629`.

[7] Y. Nesterov. "A method for solving the convex programming problem with convergence rate O(1/k^2)". In: *Proceedings of the USSR Academy of Sciences* 269 (1983), pp. 543–547.

[8] Maurice Sion. "On general minimax theorems." In: *Pacific J. Math.* 8.1 (1958), pp. 171–176. URL: `https://projecteuclid.org:443/euclid.pjm/1103040253`.

[9] Kiran K Thekumparampil et al. "Efficient algorithms for smooth minimax optimization". In: *Advances in Neural Information Processing Systems*. 2019, pp. 12680–12691.

[10] Yhli. *Que-sais je?* May 2017. URL: `http://yenhuanli.github.io/blog/2017/05/05/mirror-descent-str/`.