Efficient Algorithms for Smooth Minimax Optimisation in Non-Euclidean Space Using Bregman Divergence Framework

Nitish V. Deshpande, Sandeep K. Routray
Department of Electrical Engineering, Indian Institute of Technology Kanpur
Emails: nitishvd@iitk.ac.in, sroutray@iitk.ac.in

EE698U Term Presentation

Motivation

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.

Motivation of analysis in Non-Euclidean Space

(1) The quadratic proximity term $\frac{1}{2\eta_k}\|x - x_k\|_2^2$ is inappropriate for problems with highly inhomogeneous geometry

(2) For example, the quadratic minimization problem $\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}(x - x_0)^T Q(x - x_0)$ where $Q \succ 0$ is a diagonal matrix with high condition number. The inhomogeneous geometry leads to slow updates (since iteration complexity depends on condition number for the quadratic minimization problem) in the conventional GD algorithm.

(3) For the probability simplex problem, Euclidean distance is in general not recommended for measuring the distance between probability vectors

(4) 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 $\frac{1}{2\eta_k}\|x - x_k\|_2^2$ by a class of general distance-like metric known as the Bregman divergence

Prior work

(1) Work on Smooth Minimax Optimisation:[1] $\tilde{\mathcal{O}}(1/k^2)$ convergence rate for smooth, strongly-convex – concave problems, improving upon the previous best known rate of $\mathcal{O}(1/k)$

(2) Work on Non-Euclidean analysis:

- **(1)** General Norm[2] They analyse Nesterov's accelerated gradient descent using general norm for the unconstrained case. They develop a potential function based framework for proving convergence rate.
- **(2)** Relative strong convexity and smoothness:[3] They 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.
- **(3)** Riemannian space:[4] Proposed a Riemannian counterpart to Nesterov's AGD.

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

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

[3] 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.

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

Problem Formulation

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

Assumptions

(A1) $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}$ where $\mathcal{V}$ is a normed vector space with an arbitrary norm $\|.\|$ on the underlying space

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

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

(A4) $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is L-smooth

$$\max \left\{ \left\| \nabla_x g(x,y) - \nabla_x g(x',y') \right\|_* , \left\| \nabla_y g(x,y) - \nabla_y g(x',y') \right\|_* \right\} \leq L \left(\|x - x'\| + \|y - y'\| \right)$$

Our goal To find a ϵ-primal-dual pair $(\hat{x}, \hat{y})$ defined as: $(\hat{x}, \hat{y})$ is an ϵ-primal-dual pair of g if the primal-dual gap is less than ϵ: $\max_{y \in \mathcal{Y}} g(\hat{x}, y) - \min_{x \in \mathcal{X}} g(x, \hat{y}) \leq \epsilon$

Contributions

(1) Using potential function based framework[5] and Bregman divergence framework, proved $\mathcal{O}(1/k^2)$ convergence rate for Nesterov's AGD for the *general norm* and *constrained case*.

(2) Proposed Generalized Conceptual Dual Implicit Accelerated Gradient Descent (GC-DIAG) which is adapted from the Conceptual Dual Implicit Accelerated Gradient (C-DIAG)[6] and proved $\mathcal{O}(1/k^2)$ convergence rate for the primal dual gap.

(3) Proved $\mathcal{O}(\frac{1}{k^4})$ convergence rate using Nesterov's AGD and restarting strategy which is an improvement over $\mathcal{O}(\frac{1}{k})$ for smooth and strongly convex functions with respect to an arbitrary norm.

	Smooth and convex	Smooth and strongly convex
Mirror descent	$\mathcal{O}(\frac{1}{\sqrt{k}})$	$\mathcal{O}(\frac{1}{k})$[7]
Nesterov's AGD	$\mathcal{O}(\frac{1}{k^2})$[8]	$\mathcal{O}(\frac{1}{k^4})$

Table: Comparison of oracle complexities with arbitrary norm

[5] Bansal and Gupta, "Potential-function proofs for first-order methods".

[6] Thekumparampil et al., "Efficient algorithms for smooth minimax optimization".

[7] Yhli. *Minimizing a Strongly Convex Function by Mirror Descent*. 2017. URL: http://yenhuanli.github.io/blog/2017/05/05/mirror-descent-str/.

[8] Y. Nesterov. "A method for solving the convex programming problem with convergence rate O(1/k^2)".

Nesterov's accelerated gradient ascent with general norm

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 $y_0$ and $z_0$
Output: $y_K$
for $k = 0, 1, ..., K$ do
    $w_k \leftarrow (1 - \tau_k)y_k + \tau_k z_k$    (1)
    $y_{k+1} \leftarrow \arg\min_{y \in \mathcal{Y}} \left\{ -\langle \nabla h(w_k), y - w_k \rangle + \frac{\beta}{2} \|y - w_k\|^2 \right\}$    (2)
    $z_{k+1} \leftarrow \arg\min_{z \in \mathcal{Y}} \{ -\eta_k \langle \nabla h(w_k), z \rangle + D_\psi(z\|z_k) \}$    (3)
end for
```

For Euclidean norm (2) becomes $y_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}(\omega_k + \frac{1}{\beta}\nabla h(w_k))$ and (3) becomes $z_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}(z_k + \eta_k \nabla h(w_k))$

A major hurdle in general norm case: We cannot use properties of Projection operator like Non-expansiveness

Analysis using Potential function

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

Our goal:

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

Lemma

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

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

Analysis using Potential function

$$\Phi(k+1) - \Phi(k) = (k+1)(k+2)\underbrace{\left(h(\mathrm{w}_k) - h(\mathrm{y}_{k+1})\right)}_{(a)}$$

$$\underbrace{-k(k+1)(h(\mathrm{w}_k) - h(\mathrm{y}_k)) + 2(k+1)(h(\mathrm{y}) - h(\mathrm{w}_k))}_{(b)} + \frac{4\beta}{\mu_\psi}\underbrace{\left(D_\psi(\mathrm{y}\|\mathrm{z}_{k+1}) - D_\psi(\mathrm{y}\|\mathrm{z}_k)\right)}_{(c)} \tag{4}$$

For bounding (a) and (b), we used:

(1) The fact that $(-h(\mathrm{x}))$ is L-smooth and the choice of $\beta > L$.

(2) Concavity of $h(\mathrm{x})$ and the choice of $\tau_k = \frac{2}{k+2}$.

Analysis using Potential function

$$\begin{aligned}
& D_\psi(\mathrm{y}\|\mathrm{z}_{k+1}) - D_\psi(\mathrm{y}\|\mathrm{z}_k) \\
& = (\psi(\mathrm{y}) - \psi(\mathrm{z}_{k+1}) - \langle \nabla\psi(\mathrm{z}_{k+1}), \mathrm{y} - \mathrm{z}_{k+1}\rangle) - (\psi(\mathrm{y}) - \psi(\mathrm{z}_k) - \langle \nabla\psi(\mathrm{z}_k), \mathrm{y} - \mathrm{z}_k\rangle) \\
& = \psi(\mathrm{z}_k) - \psi(\mathrm{z}_{k+1}) + \langle \nabla\psi(\mathrm{z}_k), \mathrm{z}_{k+1} - \mathrm{z}_k\rangle + \langle \nabla\psi(\mathrm{z}_{k+1}) - \nabla\psi(\mathrm{z}_k), \mathrm{z}_{k+1} - \mathrm{y}\rangle \\
& \leq -\frac{\mu_\psi}{2} \|\mathrm{z}_{k+1} - \mathrm{z}_k\|^2 + \underbrace{\langle \nabla\psi(\mathrm{z}_{k+1}) - \nabla\psi(\mathrm{z}_k), \mathrm{z}_{k+1} - \mathrm{y}\rangle}_{(d)} \qquad (5)
\end{aligned}$$

Inequality is due to μ_ψ-strongly convex function $\psi(.)$
From the update in (3) in Algorithm 1, we write the optimality condition as

$$\left\langle (-\eta_k \nabla h(\mathrm{w}_k) + \nabla_\mathrm{z} D_\psi(\mathrm{z}\|\mathrm{z}_k))\Big|_{\mathrm{z}=\mathrm{z}_{k+1}}, \mathrm{y} - \mathrm{z}_{k+1} \right\rangle \geq 0, \quad \forall \mathrm{y} \in \mathcal{Y}. \qquad (6)$$

From definition of Bregman divergence, $\nabla_\mathrm{z} D_\psi(\mathrm{z}\|\mathrm{z}_k)\Big|_{\mathrm{z}=\mathrm{z}_{k+1}} = \nabla\psi(\mathrm{z}_{k+1}) - \nabla\psi(\mathrm{z}_k)$.
Hence, the term (d) in Equation (5) can be bounded as

$$\langle \nabla\psi(\mathrm{z}_{k+1}) - \nabla\psi(\mathrm{z}_k), \mathrm{z}_{k+1} - \mathrm{y}\rangle \leq \langle \eta_k \nabla h(\mathrm{w}_k), \mathrm{z}_{k+1} - \mathrm{y}\rangle . \qquad (7)$$

Hence,

$$D_\psi(\mathrm{y}\|\mathrm{z}_{k+1}) - D_\psi(\mathrm{y}\|\mathrm{z}_k) \leq -\frac{\mu_\psi}{2} \|\mathrm{z}_{k+1} - \mathrm{z}_k\|^2 + \langle \eta_k \nabla h(\mathrm{w}_k), \mathrm{z}_{k+1} - \mathrm{y}\rangle . \qquad (8)$$

Analysis using Potential function

$$\Phi(k+1)-\Phi(k)=(k+1)(k+2)(h(\mathrm{w}_k)-h(\mathrm{y}_{k+1}))$$
$$-k(k+1)(h(\mathrm{w}_k)-h(\mathrm{y}_k))+2(k+1)(h(\mathrm{y})-h(\mathrm{w}_k))+\frac{4\beta}{\mu_\psi}(D_\psi(\mathrm{y}\|\mathrm{z}_{k+1})-D_\psi(\mathrm{y}\|\mathrm{z}_k))$$

$$h(\mathrm{y}_{k+1})-h(\mathrm{w}_k)\geq\tau_k\langle\nabla h(\mathrm{w}_k),\mathrm{z}_{k+1}-\mathrm{z}_k\rangle-\frac{\beta}{2}\tau_k^2\|\mathrm{z}_{k+1}-\mathrm{z}_k\|^2 \tag{9}$$

$$-k(k+1)(h(\mathrm{w}_k)-h(\mathrm{y}_k))+2(k+1)(h(\mathrm{y})-h(\mathrm{w}_k))\leq 2(k+1)\langle\nabla h(\mathrm{w}_k),\mathrm{y}-\mathrm{z}_k\rangle, \tag{10}$$

$$D_\psi(\mathrm{y}\|\mathrm{z}_{k+1})-D_\psi(\mathrm{y}\|\mathrm{z}_k)\leq-\frac{\mu_\psi}{2}\|\mathrm{z}_{k+1}-\mathrm{z}_k\|^2+\langle\eta_k\nabla h(\mathrm{w}_k),\mathrm{z}_{k+1}-\mathrm{y}\rangle. \tag{11}$$

$$\Phi(k+1)-\Phi(k)\leq(k+1)(k+2)\left(-\tau_k\langle\nabla h(\mathrm{w}_k),\mathrm{z}_{k+1}-\mathrm{z}_k\rangle+\frac{\beta}{2}\tau_k^2\|\mathrm{z}_{k+1}-\mathrm{z}_k\|^2\right)$$
$$+2(k+1)\langle\nabla h(\mathrm{w}_k),\mathrm{y}-\mathrm{z}_k\rangle+\frac{4\beta}{\mu_\psi}\left(-\frac{\mu_\psi}{2}\|\mathrm{z}_{k+1}-\mathrm{z}_k\|^2+\langle\eta_k\nabla h(\mathrm{w}_k),\mathrm{z}_{k+1}-\mathrm{y}\rangle\right)$$
$$\leq 2\beta\|\mathrm{z}_{k+1}-\mathrm{z}_k\|^2\left(\frac{k+1}{k+2}-1\right)+(-2(k+1)+\frac{4\beta}{\mu_\psi}\eta_k)\langle\nabla h(\mathrm{w}_k),\mathrm{z}_{k+1}-\mathrm{y}\rangle\overset{(7)}{\leq}0, \tag{12}$$

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

Generalized Conceptual Dual Implicit Accelerated Gradient Descent (GC-DIAG)

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

```
    Input: $g$, $D_\psi, \mu_\psi$, $L$, $\sigma$, $x_0$, $y_0$, $K$,
    Output: $\bar{x}_K$ $y_K$
Set $\beta > L$, $z_0 \leftarrow y_0$
for $k = 0, 1, ..., K$ do
    $\tau_k \leftarrow \frac{2}{k+2}$, $\eta_k \leftarrow \frac{k+1}{2\beta}\mu_\psi$, $w_k \leftarrow (1 - \tau_k) y_k + \tau_k z_k$
    Choose $x_{k+1}$, $y_{k+1}$, ensuring:
        $g(x_{k+1}, y_{k+1}) = \min_x g(x, y_{k+1}),$
        $y_{k+1} \leftarrow \arg\min_{y \in \mathcal{Y}} \left\{ - \langle \nabla_y g(x_{k+1}, w_k), y - w_k \rangle + \frac{\beta}{2} \|y - w_k\|^2 \right\}$
    $z_{k+1} \leftarrow \arg\min_{z \in \mathcal{Y}} \{-\eta_k \langle \nabla_y g(x_{k+1}, w_k), z \rangle + D_\psi(z \| z_k)\}$,
    $\bar{x}_{k+1} \leftarrow \frac{2}{(k+1)(k+2)} \sum_{i=1}^{k+1} i x_i$
end for
return $\bar{x}_K, y_K$
```

Convergence analysis of GC-DIAG

$$\begin{aligned}
&\underbrace{(k+1)\,(k+2)\,(g(\mathrm{x}_{k+1},\mathrm{y}) - g(\mathrm{x}_{k+1},\mathrm{y}_{k+1})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathrm{y}\|\mathrm{z}_{k+1})}_{\Phi_{k+1}} \\
&\overset{(1)}{\leq} (k)\,(k+1)\,(g(\mathrm{x}_{k+1},\mathrm{y}) - g(\mathrm{x}_{k+1},\mathrm{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathrm{y}\|\mathrm{z}_k) \\
&\overset{(2)}{=} (k)\,(k+1)\,(g(\mathrm{x}_{k+1},\mathrm{y}) - g(\mathrm{x}_k,\mathrm{y})) + (k)\,(k+1)\,(g(\mathrm{x}_k,\mathrm{y}) - g(\mathrm{x}_{k+1},\mathrm{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathrm{y}\|\mathrm{z}_k) \\
&\overset{(3)}{\leq} (k)\,(k+1)\,(g(\mathrm{x}_{k+1},\mathrm{y}) - g(\mathrm{x}_k,\mathrm{y})) + \underbrace{(k)\,(k+1)\,(g(\mathrm{x}_k,\mathrm{y}) - g(\mathrm{x}_k,\mathrm{y}_k)) + \frac{4\beta}{\mu_\psi} D_\psi(\mathrm{y}\|\mathrm{z}_k)}_{\Phi_k}
\end{aligned}$$

$$\overset{(4)}{\leq} (k)\,(k+1)\,g(\mathrm{x}_{k+1},\mathrm{y}) - \sum_{i=1}^{k}(2i)(g(\mathrm{x}_i,\mathrm{y})) + \frac{4\beta}{\mu_\psi} D_\psi(\mathrm{y}\|\mathrm{z}_0), \tag{13}$$

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

Convergence analysis of GC-DIAG

Now, we write

$$(k+1)\,(k+2)\left(g(x_{k+1},y) - g(x_{k+1},y_{k+1})\right) + \frac{4\beta}{\mu_\psi} D_\psi(y\|z_{k+1}) \tag{14}$$

$$\le (k)\,(k+1)\,g(x_{k+1},y) - \sum_{i=1}^{k}(2i)(g(x_i,y)) + \frac{4\beta}{\mu_\psi} D_\psi(y\|z_0). \tag{15}$$

Rearranging the terms, we get

$$\sum_{i=1}^{k+1}(2i)(g(x_i,y)) - (k+1)(k+2)g(x_{k+1},y_{k+1}) \le \frac{4\beta}{\mu_\psi} D_\psi(y\|z_0) - \frac{4\beta}{\mu_\psi} D_\psi(y\|z_{k+1}) \tag{16}$$

$$\sum_{i=1}^{k+1}(2i)(g(x_i,y)) - (k+1)(k+2)g(x,y_{k+1}) \overset{(5)}{\le} \frac{4\beta}{\mu_\psi} D_\psi(y\|z_0) \tag{17}$$

$$g(\bar{x}_{k+1},y)) - g(x,y_{k+1}) \overset{(6)}{\le} \frac{4\beta}{\mu_\psi(k+1)(k+2)} D_\psi(y\|z_0) \tag{18}$$

$$\max_{y\in\mathcal{Y}} g(\bar{x}_{k+1},y)) - \min_{x\in\mathcal{X}} g(x,y_{k+1}) \overset{(7)}{\le} \frac{4\beta}{\mu_\psi(k+1)(k+2)} D_\psi(y\|z_0) \tag{19}$$

where inequality (5) follows from the fact that $g(x,y_{k+1}) \ge g(x_{k+1},y_{k+1}) \forall x \in \mathcal{X}$. Inequality (6) follows by defining a convex combination of $\bar{x}_{k+1} = \frac{1}{(k+1)(k+2)} \sum_{i=1}^{k+1}(2i)x_i$ and from the fact that $g(.,y)$ is convex for every y.

Generalized Dual Implicit Accelerated Gradient Descent (G-DIAG)

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

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

Imp Step subroutine in G-DIAG

Algorithm 4 Imp Step subroutine in G-DIAG

```
Imp − STEP($g, L, \sigma, x_0, w, \beta, \epsilon_{\text{step}}^{(k+1)}$):
    Set $\epsilon_{\text{mp}} \leftarrow ?$, $R \leftarrow ?$, $\epsilon_{\text{agd}} \leftarrow ?$, $y_0 \leftarrow w$
    for $r = 0, 1, ..., R$ do
        Starting at $x_0$ use generalized AGD (Algorithm 1 with $-g(., y_r)$) to compute $x_r$ such that:
            $g(\hat{x}_r, y_r) \leq \min_x g(x, y_r) + \epsilon_{\text{agd}}$,    (20)
        $y_{k+1} \leftarrow \arg\min_{y \in \mathcal{Y}} \left\{ - \langle \nabla_y g(\hat{x}_r, w), y - w \rangle + \frac{\beta}{2} \|y - w\|^2 \right\}$
    end for
    return $\hat{x}_R, y_{R+1}$
```

In the $\mathtt{Imp} - \mathtt{Step}$ of the original DIAG algorithm,[9] AGD (with ℓ^2 norm) is used to efficiently calculate (in logarithmic number of steps) an estimate for $x_r = \arg\min_{x \in \mathcal{X}} g(x, y_r) + \epsilon_{\text{agd}}$. The proof uses the guarantee on AGD for strongly convex case.[10]

[9] Thekumparampil et al., "Efficient algorithms for smooth minimax optimization".

[10] Bansal and Gupta, "Potential-function proofs for first-order methods", Equation (5.68).

Challenges faced in arbitrary norm case

We faced the the following issues

(A) Analyzing AGD for strongly convex case in arbitrary norm

(1) 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.

(2) It may happen that, for arbitrary norm, finding an ϵ_{agd} estimate x_r would not be possible in linear time.

(3) As an example, we found a blog[11] that discusses the case of mirror descent for the strongly convex case with respect to a general norm.

(4) 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.

(B) Proving that the $\mathtt{Imp} - \mathtt{Step}$ convergence: Let $\mathsf{x}^*(\mathsf{y}) = \arg\min_{\mathsf{x}\in\mathcal{X}} g(\mathsf{x},\mathsf{y})$, then for proving that there exists a fixed point of the iterations of the $\mathtt{Imp} - \mathtt{Step}$, we require to show $\mathsf{y}^+ = \arg\min_{\mathsf{y}\in\mathcal{Y}} \left\{-\left\langle \nabla_{\mathsf{y}} g(\mathsf{x}^*(\mathsf{y}),\mathsf{w}), \mathsf{y}-\mathsf{w}\right\rangle + \frac{\beta}{2}\left\|\mathsf{y}-\mathsf{w}\right\|^2\right\}$ is a contraction, that is $\left\|\mathsf{y}_1^+ - \mathsf{y}_2^+\right\| \leq \alpha \left\|\mathsf{y}_1 - \mathsf{y}_2\right\|$, $\alpha \leq 1$.

[11] Yhli, *Minimizing a Strongly Convex Function by Mirror Descent*.

Convergence analysis for generalized AGD with strong concavity

(1) In case of mirror descent, the restarting strategy discussed in[12] 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).

(2) We try to come up with the convergence rate for generalized AGD with strong concavity following a similar procedure.

(3) Using Lemma 7, the oracle complexity for generalized AGD can be found by carrying out a telescoping sum.

Theorem

Suppose $h(.)$ is a L-smooth function and the parameters of Algorithm 1 are chosen as per Lemma 7, then the following holds

$$h(\mathrm{y}^*) - h(\mathrm{y}_T) \leq \frac{4\beta}{\mu_\psi} \cdot \frac{D_\psi(\mathrm{y}^*||\mathrm{y}_0)}{T(T+1)} \tag{21}$$

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

$$h(\mathrm{y}^*) - h(\mathrm{y}_T) \leq \frac{4\beta}{\mu_\psi} \cdot \frac{\Omega}{T(T+1)} \tag{22}$$

[12] Yhli, *Minimizing a Strongly Convex Function by Mirror Descent*.

Convergence analysis for generalized AGD with strong concavity

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 u, define $\psi_{R,\mathbf{u}}(\mathrm{y}) := \psi(R^{-1}(\mathrm{y} - \mathrm{u}))$. Let $D_{\psi,R,\mathbf{z}}(.\|.)$ denote the corresponding Bregman divergence.

Corollary

Suppose

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

If we apply Algorithm 1 with $\eta_k = \frac{(k+1)}{2R_\mathbf{0}\beta}$, $\tau_k = \frac{2}{k+2}$, learning rate $\frac{1}{R_\mathbf{0}\beta}$ for some $\beta > L$ and Bregman divergence $D_{\psi,R,\mathbf{y_0}}(.||.)$ for T iterations. Then, the following bounds hold

$$h(\mathrm{y}^*) - h(\mathrm{y}_T(R_0, \mathrm{y}_0)) \leq \frac{4R_0\beta}{\mu_\psi} \cdot \frac{\Omega}{T(T+1)} \tag{23}$$

$$\|\mathrm{y}^* - \mathrm{y}_T(R_0, \mathrm{y}_0)\|^2 \leq \frac{8R_0\beta}{\mu_\psi\mu} \cdot \frac{\Omega}{T(T+1)} \tag{24}$$

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

Convergence analysis for generalized AGD with strong concavity

Proof.

Consider the norm $\|.\|_{R_{\mathbf{0}}} := R_0^{-1}\|.\|$. Note that $h(.)$ is $R_0 L$-smooth and $\psi_{R,\mathbf{u}}(.)$ is μ_ψ-strongly convex with respect to $\|.\|_{R_{\mathbf{0}}}$. In this case, Ω would become

$$D_{\psi, R_{\mathbf{0}}, \mathbf{y_0}}(\mathrm{y}^* || \mathrm{y}_0) = D_\psi(R_0^{-1}(\mathrm{y}^* - \mathrm{y}_0)||0) = \max\left\{ D_\psi(\mathrm{v}||0) | \, \|\mathrm{v}\| \leq 1, \mathrm{v} \in \mathbb{R}^d \right\} := \Omega$$

The bound in (23) follows directly from Theorem 17. The bound in Equation (24) is obtained from (23) using strong-concavity of $h(.)$ and the optimality of y^*.

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

□

Convergence analysis for generalized AGD with strong concavity

The error bounds given by Equations (23) (24) depend on R_0 with smaller R_0 giving smaller error bounds. Also, the bound of the distance between y^* and y_T (24) is strictly decreasing with iterations T. These observations can be used to design the following restarting strategy.

Restarting strategy

(1) Set $\mathrm{y}_0 \in \mathcal{Y}$, $l = 0$

(2) Set T_l such that $\|\mathrm{y}^* - \mathrm{y}_{T_l}(R_l, \mathrm{y}_l)\|^2 \leq 2^{-1} R_l^2$.

(3) Compute $\mathrm{y}_{l+1} = \mathrm{y}_{T_l}(R_l, \mathrm{y}_l)$ using generalized AGD as per Corollary 18.

(4) Set $R_{l+1}^2 = 2^{-1} R_l^2$, $l = l + 1$. Go to step 2.

By Corollary 18, it suffices to choose T_l such that

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

$$\implies T_l = \left\lceil \sqrt{\frac{16\beta\Omega}{R_l \mu_\psi \mu}} \right\rceil \tag{25}$$

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

Convergence analysis for generalized AGD with strong concavity

Proposition

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

$$h(\mathrm{y}^*) - h(\mathrm{y}_L) \leq 2^{-(0.5M_L+1)}\mu R_0^2, \quad \text{for } L \leq L^* \tag{26}$$

$$\|\mathrm{y}^* - \mathrm{y}_L\|^2 \leq 2^{-0.5M_L} R_0^2, \quad \text{for } L \leq L^* \tag{27}$$

and

$$h(\mathrm{y}^*) - h(\mathrm{y}_L) \leq \frac{1024\beta^2\Omega^2}{\mu_\psi^2\mu M_L^4}, \quad \text{for } L > L^* \tag{28}$$

$$\|\mathrm{y}^* - \mathrm{y}_L\|^2 \leq \frac{2048\beta^2\Omega^2}{\mu_\psi^2\mu^2 M_L^4}, \quad \text{for } L > L^* \tag{29}$$

Convergence analysis for generalized AGD with strong concavity

Proof.

Using the proposed restarting strategy, it follows from Corollary 18 that

$$h(y^*) - h(y_L) \leq 2^{-(L+1)} \mu R_0^2 \tag{30}$$

$$\|y^* - y_L\|^2 \leq 2^{-L} R_0^2 \tag{31}$$

By the choice of T_l given in Equation (25), 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

$$M_L \leq 2L, \quad \text{for } L \leq L^* \tag{32}$$

$$M_L \leq \sqrt{\frac{32\beta\Omega}{R_0 \mu_\psi \mu}} 2^{(L+1)/4}, \quad \text{for } L > L^* \tag{33}$$

The bounds given in Equations (26)-(29) follow by eliminating L from Equations (30), (31) using the relations in Equations (32), (33). □

Conclusion and future work

(1) 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

(2) 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 $\mathcal{O}(\frac{1}{k^{\mathbf{4}}})$ using the restarting strategy

(3) 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.

Thank You