Unifying mirror descent and dual averaging

Anatoli Juditsky **Joon Kwon** Éric Moulines

INRAE & AgroParisTech

6 octobre 2020
Séminaire Palaisien

Summary

- Recall the mirror descent and dual averaging algorithms,

  which are extensions of gradient descent and discuss similarities and differences, and give an informal comparison of iterates behavior.

- Define a unifying family of algorithms

  and present key tools for its analysis.

- GoLD: a new algorithm for constrained optimization

  that belongs to the unifying family. The algorithm is defined with the idea of combining the avantages of mirror descent and dual averaging.

- Present the adaptation for solving variational inequalities

  which gives a family of algorithms unifiying both mirror prox and dual extrapolation.

Unconstained optimization: Gradient descent

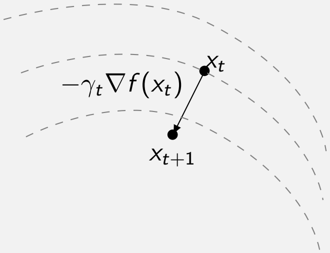
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, and assume that there exists a minimizer $x_* \in \mathbb{R}^d$.

$$f(x_*) = \min_{x \in \mathbb{R}^d} f(x)$$

Gradient descent

$$x_1 \in \mathbb{R}^d$$

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t)$$



Unconstained optimization: Gradient descent

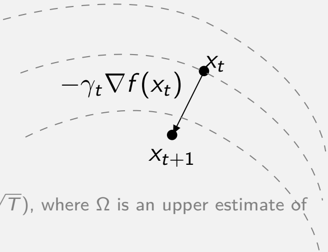
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, and assume that there exists a minimizer $x_* \in \mathbb{R}^d$.

$$f(x_*) = \min_{x \in \mathbb{R}^d} f(x)$$

Gradient descent

$$x_1 \in \mathbb{R}^d$$

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t)$$



- If f is M-Lipschitz, the choice $\gamma_t = \Omega/(M\sqrt{T})$, where Ω is an upper estimate of $\|x_* - x_1\|_2$, guarantees

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x_*) \leqslant \frac{\Omega M}{\sqrt{T}}.$$

- If ∇f is L-Lipschitz, $\gamma_t = 1/L$ guarantees

$$f(x_{T+1}) - f(x_*) \leqslant \frac{L\,\|x_1 - x_*\|_2^2}{2T}.$$

Unconstrained optimization: mirror descent

When the assumptions are satisfied with respect to a norm different from the Euclidean norm, a different algorithm may give better guarantees.

- A «proximal» rewriting of gradient descent

$$
\begin{aligned}
x_{t+1} &= x_t - \gamma_t \nabla f(x_t) \\
&= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}.
\end{aligned}
$$

Unconstrained optimization: mirror descent

When the assumptions are satisfied with respect to a norm different from the Euclidean norm, a different algorithm may give better guarantees.

- A «proximal» rewriting of gradient descent

$$\begin{aligned} x_{t+1} &= x_t - \gamma_t \nabla f(x_t) \\ &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}. \end{aligned}$$

- A different *geometry* gives mirror descent (Nemirovsky & Yudin, 1983; Beck & Teboule, 2003):

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\gamma_t} D_F(x, x_t) \right\},$$

where the Bregman divergence is defined by

$$D_F(x', x) := F(x) - F(x') - \left\langle \nabla F(x), x - x' \right\rangle,$$

is associated with a differentiable mirror map F.

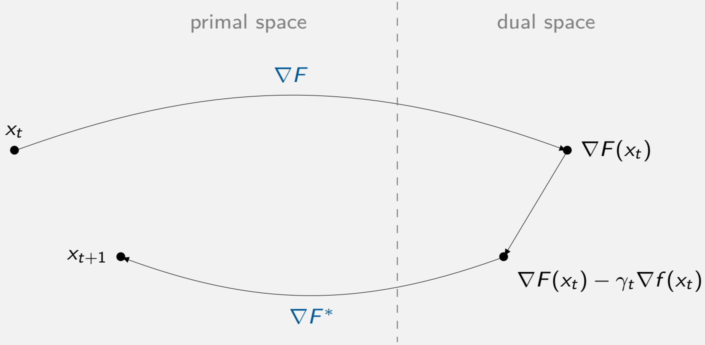
Unconstrained optimization: mirror descent

- A «primal–dual» rewriting of mirror descent

$$x_{t+1} = \nabla F^*(\nabla F(x_t) - \gamma_t \nabla f(x_t)),$$

where F^* is the Legendre–Fenchel transform of F:

$$F^*(y) = \max_{x \in \mathbb{R}^d} \{\langle y, x\rangle - F(x)\}, \quad y \in \mathbb{R}^d.$$



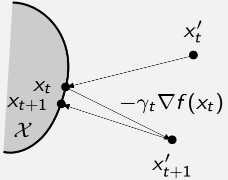
Constrained optimization: use the Euclidean projection

Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and convex. We assume that f admits a minimizer x_* on $\mathcal{X}$.

$$f(x_*) = \min_{x \in \mathcal{X}} f(x).$$

- Projected gradient descent (Goldstein, 1964; Levitin & Polyak, 1966)

$$x'_{t+1} = x_t - \gamma_t \nabla f(x_t)$$
$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x'_{t+1})$$



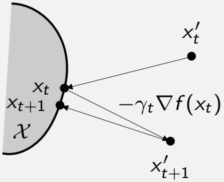
Constrained optimization: use the Euclidean projection

Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and convex. We assume that f admits a minimizer x_* on $\mathcal{X}$.

$$f(x_*) = \min_{x \in \mathcal{X}} f(x).$$

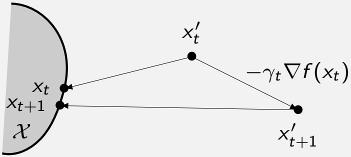
- **Projected gradient descent** (Goldstein, 1964; Levitin & Polyak, 1966)

$$x'_{t+1} = x_t - \gamma_t \nabla f(x_t)$$
$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x'_{t+1})$$



- **Dual averaging** (Shalev-Shwartz, 2007; Nesterov, 2009; Xiao, 2010)

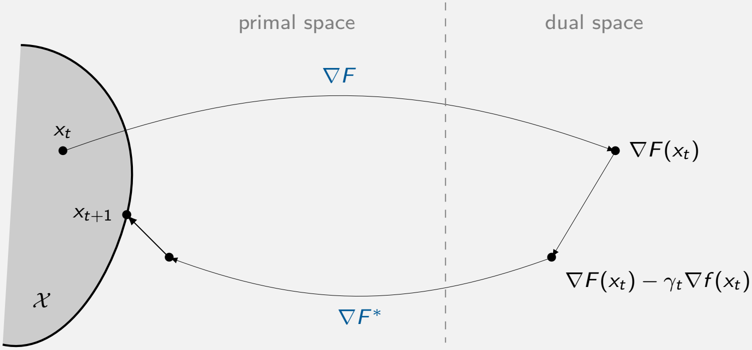
$$x'_{t+1} = x'_t - \gamma_t \nabla f(x_t)$$
$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x'_{t+1})$$



Constrained optimization: mirror descent

- Mirror descent is the extension of projected gradient descent.
- The Euclidean projection is replaced by a projection is with respect to the Bregman divergence associated with mirror map F.

$$x'_{t+1} = \nabla F^*(\nabla F(x_t) - \gamma_t \nabla f(x_t))$$
$$x_{t+1} = \arg\min_{x \in \mathcal{X}} D_F(x, x'_{t+1}).$$



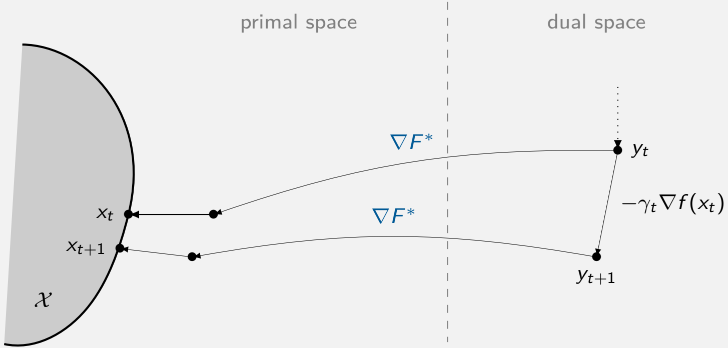
- Mirror map F must satisfy assumptions so that mirror descent iterates on $\mathcal{X}$ are well-defined. We then say that F is compatible with $\mathcal{X}$.

Constrained optimization: dual averaging

- Dual averaging (Nesterov, 2009; Xiao, 2010) is the extension of the Euclidean algorithm which performs the gradient step from the «unprojected» point.

$$x'_{t+1} = \nabla F^* \left(y_1 - \sum_{s=1}^{t} \gamma_s \nabla f(x_s) \right).$$

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} D_F(x, x'_{t+1}).$$



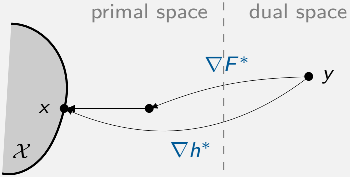
From mirror maps to regularizers

Lemma

Let F be a mirror map compatible with $\mathcal{X}$. Then,

$$\arg\min_{x \in \mathcal{X}} D_F(x, \nabla F^*(y)) = \nabla h^*(y),$$

where $h = F + I_{\mathcal{X}}$.



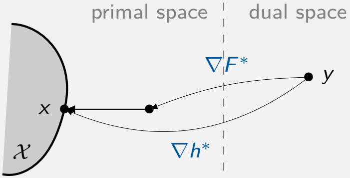
From mirror maps to regularizers

Lemma

Let F be a mirror map compatible with $\mathcal{X}$. Then,

$$\arg\min_{x \in \mathcal{X}} D_F(x, \nabla F^*(y)) = \nabla h^*(y),$$

where $h = F + I_{\mathcal{X}}$.



Definition (Regularizers)

A function $h \colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a regularizer on $\mathcal{X}$ if:

- h is convex and lowersemicontinuous,
- $\operatorname{cl} \operatorname{dom} h = \mathcal{X}$,
- $\operatorname{dom} h^* = \mathbb{R}^d$.

A mirror map F has an associated regularizer $h = F + I_{\mathcal{X}}$. The converse is not always true.

Informal comparison

Mirror descent and dual averaging coincide if

the problem is unconstrained

or if iterates lie in the interior of $\mathcal{X}$

When different, dual averaging is more conservative than mirror descent

mirror descent converges faster but for precisely chosen $(\gamma_t)_{t \geqslant 1}$

dual averaging converges slower but for much wider range of $(\gamma_t)_{t \geqslant 1}$

UMD: A new family of algorithms

Definition

- Let h be a regularizer on $\mathcal{X}$. Let $\Pi_h : \mathcal{X} \times \mathbb{R}^d \rightrightarrows \mathbb{R}^d \times \mathbb{R}^d$ defined as

$$(x, y) \in \Pi_h(y_0) \quad \Longleftrightarrow \quad \begin{cases} x = \nabla h^*(y_0) \\ y \in \partial h(x) \\ \forall x' \in \mathcal{X}, \quad \langle y - y_0, x' - x \rangle \geqslant 0. \end{cases}$$

UMD: A new family of algorithms

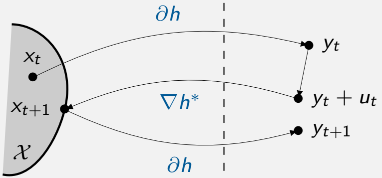
Definition

- Let h be a regularizer on $\mathcal{X}$. Let $\Pi_h : \mathcal{X} \times \mathbb{R}^d \rightrightarrows \mathbb{R}^d \times \mathbb{R}^d$ defined as

$$(x, y) \in \Pi_h(y_0) \quad \Longleftrightarrow \quad \begin{cases} x = \nabla h^*(y_0) \\ y \in \partial h(x) \\ \forall x' \in \mathcal{X}, \quad \langle y - y_0, x' - x \rangle \geqslant 0. \end{cases}$$

- Let $(u_t)_{t \geqslant 1}$ be a sequence in $\mathbb{R}^d$ (e.g. $u_t = -\gamma_t \nabla f(x_t)$). $(x_t, y_t)_{t \geqslant 1}$ is sequence of UMD iterates if

$$\forall t \geqslant 1, \quad (x_{t+1}, y_{t+1}) \in \Pi_h(y_t + u_t).$$



Proposition

Mirror descent and dual averaging are special cases of UMD.

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle \,.$$

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle .$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t), \\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t). \end{aligned}$$

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle .$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t), \\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t). \end{aligned}$$

Proof.

$$\langle u_t, x - x_{t+1} \rangle \leqslant \langle y_{t+1} - y_t, x - x_{t+1} \rangle$$

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle .$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t), \\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t). \end{aligned}$$

Proof.

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant \langle y_{t+1} - y_t, x - x_{t+1} \rangle \\ &= \langle y_{t+1}, x - x_{t+1} \rangle - \langle y_t, x - x_t \rangle + \langle y_t, x_{t+1} - x_t \rangle \end{aligned}$$

.

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle \,.$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned}\langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t), \\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t).\end{aligned}$$

Proof.

$$\begin{aligned}\langle u_t, x - x_{t+1} \rangle \leqslant \langle y_{t+1} - y_t, x - x_{t+1} \rangle & \\ = \langle y_{t+1}, x - x_{t+1} \rangle - \langle y_t, x - x_t \rangle + \langle y_t, x_{t+1} - x_t \rangle & \\ + h(x) - h(x) + h(x_t) - h(x_t) + h(x_{t+1}) - h(x_{t+1}) & \\ . &\end{aligned}$$

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle .$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned}\langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t),\\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t).\end{aligned}$$

Proof.

$$\begin{aligned}\langle u_t, x - x_{t+1} \rangle &\leqslant \langle y_{t+1} - y_t, x - x_{t+1} \rangle\\ &= \langle y_{t+1}, x - x_{t+1} \rangle - \langle y_t, x - x_t \rangle + \langle y_t, x_{t+1} - x_t \rangle\\ &\quad + h(x) - h(x) + h(x_t) - h(x_t) + h(x_{t+1}) - h(x_{t+1})\\ &= D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t).\end{aligned}$$

UMD: Analysis

Definition (Generalized Bregman divergence)

Let $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. For $x, x', y \in \mathbb{R}^n$ such that $y \in \partial h(x)$, we define

$$D_h(x', x;\ y) = h(x') - h(x') - \langle y, x' - x \rangle .$$

Lemma (Regret bound)

For $x \in \operatorname{dom} h$, and all $t \geqslant 1$,

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t), \\ \langle u_t, x - x_t \rangle &\leqslant D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) + D_{h^*}(y_t + u_t, y_t). \end{aligned}$$

Proof.

$$\begin{aligned} \langle u_t, x - x_{t+1} \rangle &\leqslant \langle y_{t+1} - y_t, x - x_{t+1} \rangle \\ &= \langle y_{t+1}, x - x_{t+1} \rangle - \langle y_t, x - x_t \rangle + \langle y_t, x_{t+1} - x_t \rangle \\ &\qquad + h(x) - h(x) + h(x_t) - h(x_t) + h(x_{t+1}) - h(x_{t+1}) \\ &= D_h(x, x_t;\ y_t) - D_h(x, x_{t+1};\ y_{t+1}) - D_h(x_{t+1}, x_t;\ y_t). \end{aligned}$$

$$\langle u_t, x_{t+1} - x_t \rangle = D_h(x_{t+1}, x_t;\ y_t) + \underbrace{D_h(x_t, x_{t+1};\ y_t + u_t)}_{= D_{h^*}(y_t + u_t;\ y_t)} .$$

Exemples of guarantees

$f : \mathbb{R}^d \to \mathbb{R}$ convex and differentiable, $f(x_*) = \min_{x\in\mathcal{X}} f(x)$.
Regularizer h is assumed to be K-strongly convex with respect to a give norm $\|\cdot\|$:

$$\forall x, x' \in \mathbb{R}^d, \quad h(\lambda x + (1-\lambda)x') \leqslant \lambda h(x) + (1-\lambda)h(x') - \frac{K\lambda(1-\lambda)}{2}\left\|x' - x\right\|^2.$$

Consider UMD iterates for minimizing f:

$$(x_{t+1}, y_{t+1}) \in \Pi_h(y_t - \gamma_t \nabla f(x_t)).$$

Theorem (Lipschitz convex optimization)

If f is M-Lipschitz with respect to $\|\cdot\|$, the choice $\gamma_t = \Omega_{\mathcal{X}} M^{-1}\sqrt{K/T}$ guarantees

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x_*) \leqslant \frac{\Omega_{\mathcal{X}} M}{\sqrt{KT}},$$

where $\Omega_{\mathcal{X}}$ is an upper-estimate of $\sqrt{2D_h(x_*, x_1;\ y_1)}$.

Theorem (Smooth convex optimization)

If ∇f is L-Lipschitz with respect to $\|\cdot\|_*$ and $\|\cdot\|$, the choice $\gamma_t = K/L$ guarantees

$$f(x_{T+1}) - f(x_*) \leqslant \frac{LD_h(x_*, x_1;\ y_1)}{KT}.$$

Guarantees can be derived for stochastic and/or non-convex settings.

Exemples of derivative algorithms

- Last-iterate convergence for Lipschitz convex optimization. Extension of (Nazin, 2018).

$$(x_{t+1}, y_{t+1}) \in \Pi_h(y_t - \gamma_t \nabla f(x_t^+)) \qquad \nu_t = \gamma_{t+1}\left(\sum_{s=1}^{t+1}\gamma_s\right)^{-1}$$

$$x_{t+1}^+ = (1-\nu_t)x_t^+ + \nu_t x_{t+1} \qquad \gamma_t = \frac{\Omega_{\mathcal{X}}}{M\sqrt{T}}.$$

guarantees

$$f(x_T^+) - f_* \leqslant \frac{\Omega_{\mathcal{X}} M}{\sqrt{T}},$$

where $\Omega_{\mathcal{X}}$ is an upper-estimate of $\sqrt{2D_h(x_*, x_1;\ y_1)}$.

- Accelerated convergence for smooth convex optimization. Extension of (Nesterov, 1983; Nesterov, 2005; Krichene et al., 2015).

$$x_t^+ = (1-\nu_t)z_t + \nu_t x_t \qquad \gamma_1 = K/L$$

$$(x_{t+1}, y_{t+1}) \in \Pi_h(y_t - \gamma_t \nabla f(x_t^+)) \qquad \gamma_{t+1} = \frac{K}{2L}\left(1 + \sqrt{1 + (2L\gamma_t/K)^2}\right)$$

$$z_{t+1} = x_t^+ + \nu_t(x_{t+1} - x_t) \qquad \nu_t = \frac{K}{L\gamma_t},$$

guarantees

$$f(z_{T+1}) - f_* \leqslant \frac{4LD_h(x_*, x_1;\ y_1)}{KT^2}.$$

GoLD: Greedy or Lazy Descent

Idea: Every k steps, compares the objective values resulting from a dual averaging iteration and a mirror descent iteration, and retains the best. For the $k-1$ other steps, performs dual averaging iterations.

Let F be a mirror map compatible with $\mathcal{X}$ and $h = F + \mathcal{I}_{\mathcal{X}}$ the associated regularizer

```
Input: Parameter k ⩾ 1, time horizon T ⩾ 1, initial dual iterate y_1 ∈ ℝ^d.
x_2 ← ∇h*(y_1 − γ_1 ∇f(x_1))
for t = 2, ..., T − 1 do
    if t ≡ 2 mod k then
        y_t^MD ← ∇F(x_t)
        y_t^DA ← y_{t−1} − γ_{t−1} ∇f(x_{t−1})
        if f(∇h*(y_t^MD − γ_t ∇f(x_t))) ⩽ f(∇h*(y_t^DA − γ_t ∇f(x_t))) then
            y_t ← y_t^MD
        end
        else
            y_t ← y_t^DA
        end
    else
        y_t ← y_{t−1} − γ_{t−1} ∇f(x_{t−1})
    end
    end
    x_{t+1} ← ∇h*(y_t − γ_t ∇f(x_t))
end
```

Algorithm 1: k-GoLD

GoLD: Greedy or Lazy Descent

Idea: The comparison looks τ step ahead. Specifically, compares the objective value resulting from

- 1 mirror descent iteration followed by $\tau - 1$ dual averaging iterations,
- τ dual averging iterations,

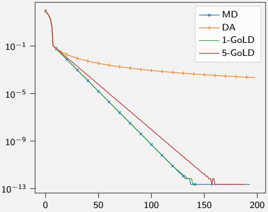
and then choose best scenario.

```
Input: Parameters k > τ ⩾ 1, time horizon T ⩾ 1, initial dual iterate y_1 ∈ ℝ^d.
x_2 ← ∇h*(y_1 − γ_1 ∇f(x_1))
for t = 2, ..., T − 1 do
    if t ≡ 2 mod k then
        y_t^MD ← ∇F(x_t)
        y_t^DA ← y_{t−1} − γ_{t−1} ∇f(x_{t−1})
        for s = 1, ..., τ do
            x_{t+s}^MD = ∇h*( y_t^MD − Σ_{s'=0}^{s−1} γ_{t+s'} ∇f(x_{t+s'}^MD) )
            x_{t+s}^DA = ∇h*( y_t^DA − Σ_{s'=0}^{s−1} γ_{t+s'} ∇f(x_{t+s'}^DA) )
        end
        if f(x_{t+τ}^MD) ⩽ f(x_{t+τ}^DA) then
            y_t ← y_t^MD
        end
        else
            y_t ← y_t^DA
        end
    end
    else
        y_t ← y_{t−1} − γ_{t−1} ∇f(x_{t−1})
    end
    x_{t+1} ← ∇h*(y_t − γ_t ∇f(x_t))
end
```

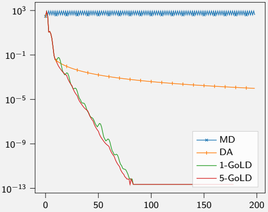
Algorithm 2: k-τ-GoLD

Numerical experiments: constrained least-squares regression

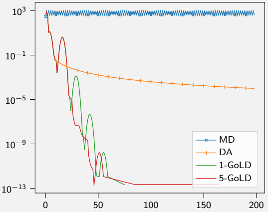
Dataset: Training sample of the BlogFeedback dataset
https://archive.ics.uci.edu/ml/datasets/BlogFeedback



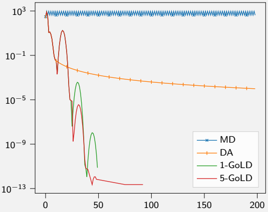
(a) $\gamma = 3 \cdot 10^{-8}$



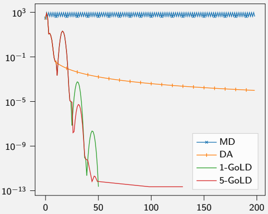
(b) $\gamma = 10^{-7}$



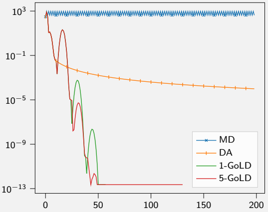
(c) $\gamma = 10^{-6}$



(d) $\gamma = 10^{-5}$



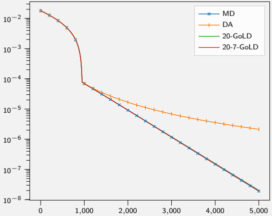
(e) $\gamma = 1$



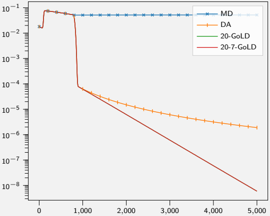
(f) $\gamma = 10^{40}$

Numerical experiments: constrained logistic regression

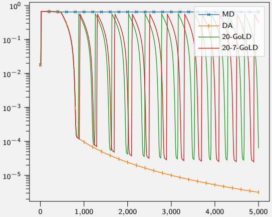
Dataset: Training sample of the Madelon dataset
https://archive.ics.uci.edu/ml/datasets/Madelon



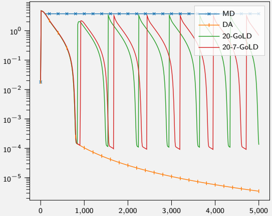
(a) $\gamma = 6 \cdot 10^{-2}$



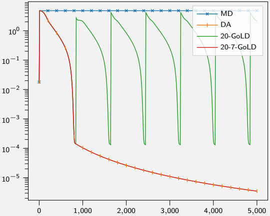
(b) $\gamma = 7 \cdot 10^{-2}$



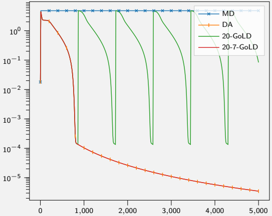
(c) $\gamma = .1$



(d) $\gamma = .3$



(e) $\gamma = 1$



(f) $\gamma = 200$

Solving variational inequalities

Let $\Phi : \mathcal{X} \to \mathbb{R}^d$ be a monotone operator:

$$\forall x, x' \in \mathcal{X}, \quad \langle \Phi(x') - \Phi(x), x' - x \rangle \geqslant 0.$$

A point $x_* \in \mathcal{X}$ is a (weak) solution if:

$$\forall x \in \mathcal{X}, \quad \langle \Phi(x), x_* - x \rangle \leqslant 0.$$

Instances include: convex optimization, convex-concave saddle-point problems, convex Nash equilibirum problems, etc.

Unified mirror prox (UMP) iterates are defined as $x_1 = \nabla h^*(y_1)$ and for $t \geqslant 1$:

$$\begin{aligned}
&w_t \in \partial h(x_t) \quad \text{such that} \quad \forall x \in \mathcal{X},\ \langle w_t - y_t, x - x_t \rangle \geqslant 0\\
&z_t = \nabla h^*(w_t - \gamma \Phi(x_t))\\
&(x_{t+1}, y_{t+1}) \in \Pi_h(y_t - \gamma \Phi(z_t)).
\end{aligned}$$

UMP contains mirror prox (Nemirovski, 2004) and dual extrapolation (Nesterov, 2007).

Theorem

If Φ is L-Lipschitz continuous with respect to $\|\cdot\|$ and $\|\cdot\|_*$:

$$\left\|\Phi(x) - \Phi(x')\right\|_* \leqslant L \left\|x - y\right\|, \quad x, x' \in \mathcal{X},$$

and h is K-strongly convex with respect $\|\cdot\|$, UMP iterates with $\gamma \leqslant K/L$ gives the following approximate weak solution

$$\forall x \in \operatorname{dom} h, \quad \left\langle \Phi(x), \frac{1}{T}\sum_{t=1}^{T} z_t - x \right\rangle \leqslant \frac{D_h(x, x_1;\ y_1)}{\gamma T}.$$

Discussion and perspectives

- Consider iterations other than mirror descent or dual averaging
  The GoLD algorithm only uses mirror descent or dual averaging iterations. We could consider algorithm which uses other iterations allowed by UMD.
- Further study of algorithms for variational inequalities
  UMP contains a simple algorithm other than mirror prox and dual extrapolation, that is to be studied.
- Extension to composite problems
  where the objection function writes $f + g$ where f is smooth and g is proxable
- Extension to time-varying regularizers
  and application to adaptive algorithms like AdaGrad, Adam, etc.