# ISyE8813
# Mirror Descent Methods

Renato DC Monteiro

January 24, 2024

Mirror Descent Method (MDM) is similar to SM except that it is based on a Bregman distance instead of the Euclidean one.

It is assumed below that $\langle \cdot, \cdot \rangle$ is an arbitrary inner product in $\mathbb{R}^n$ and that $\| \cdot \|$ is an arbitrary norm in $\mathbb{R}^n$, i.e., it is not necessarily the one associated with the inner product $\langle \cdot, \cdot \rangle$.

The dual norm $\| \cdot \|_*$ associated with $\| \cdot \|$ is then defined as

$$\|p\|_* = \max\{\langle p, x\rangle : \|x\| \leq 1\} \quad \forall p \in \mathbb{R}^n.$$

It can be easily seen that

$$\langle p, x\rangle \leq \|p\|_* \|x\| \quad \forall x, p \in \mathbb{R}^n.$$

## 1 Bregman distances

**Definition 1.1** $w \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ *is called a* **distance generating function** *if*

*(i)* $\mathrm{int}\,(\mathrm{dom}\, w) = \{x \in \mathbb{R}^n : \partial w(x) \neq \emptyset\}$*;*

*(ii)* $w$ *is continuously differentiable on* $\mathrm{int}\,(\mathrm{dom}\, w)$*.*

Define

$$W^0 := \mathrm{int}\,(\mathrm{dom}\, w), \quad W = \mathrm{dom}\, w$$

Function $w$ as in Definition 1.1 induces the Bregman distance $dw : \mathbb{R}^n \times W^0 \to \mathbb{R}$ defined for every $(x', x) \in \mathbb{R}^n \times W^0$ as

$$\begin{aligned}(dw)(x'; x) &:= w(x') - \ell_w(x'; x) \\ &= w(x') - [w(x) + \langle \nabla w(x), x' - x\rangle]\end{aligned}$$

**Remark:** For every $(x', x) \in \mathbb{R}^n \times W^0$, have

$$(dw)(x'; x) \geq 0$$

For simplicity, for every $x \in W^0$, the function $(dw)(\cdot; x)$ will be denoted by $(dw)_x$ so that

$$(dw)_x(x') = (dw)(x'; x) \quad \forall x' \in \mathbb{R}^n.$$

**Remark:** It is well known that for any $w \in \mathrm{Conv}(\mathbb{R}^n)$, we have

$$\emptyset \neq \mathrm{ri}\,(\mathrm{dom}\, w) \subset \{x \in \mathbb{R}^n : \partial w(x) \neq \emptyset\}$$

This fact and Definition 1.1(i) imply that $W^0 \neq \emptyset$.

**Exercise:** Show that conditions (i) and (ii) of Definition 1.1 are equivalent to the condition that $w$ is differentiable over the set $\{x \in \mathbb{R}^n : \partial w(x) \neq \emptyset\}$

**Lemma 1.2** *For every* $x, x' \in W^0$ *and* $u \in \operatorname{dom} w$*, we have:*

$$\nabla(dw)_x(x') = -\nabla(dw)_{x'}(x) = \nabla w(x') - \nabla w(x)$$

$$(dw)_{x'}(u) - (dw)_x(u) = \langle \nabla w(x) - \nabla w(x'), u - x \rangle + (dw)_{x'}(x)$$

**Proof**: Exercise.

**Definition 1.3** *Let* $\nu > 0$ *and convex set* $X \neq \emptyset$ *be given. A distance generating function* $w$ *is called a* $\nu$**-distance generating function for** $X$ *if*

*i)* $\operatorname{ri} X \subset W^0$ *and* $X \subset W$*;*

*ii)* $w$ *is* $\nu$*-strongly convex on* $X$*;*

**Remark:** For every $(x', x) \in \mathbb{R}^n \times W^0$, have

$$(dw)(x'; x) \geq \frac{\nu}{2}\|x' - x\|^2$$

Here are some classical and useful examples of distance generating functions.

**Example 1:** If $\|\cdot\|$ is the inner product norm, then $w(\cdot) = \|\cdot\|^2/2$ is a 1-distance generating function for any convex set $X$ and

$$dw_x(x') = \frac{1}{2}\|x' - x\|^2 \quad \forall x, x' \in \mathbb{R}^n$$

**Example 2:** If $\|\cdot\| = \|\cdot\|_1$ where

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \quad \forall x \in \mathbb{R}^n,$$

then function $w : \mathbb{R}_+^n \to \mathbb{R}$ defined as

$$w(x) = \sum_{i=1}^{n} x_i \log x_i$$

is a 1-distance generating function for

$$\Delta_n := \{x \in \mathbb{R}_+^n : \langle e, x \rangle = 1\}$$

where $e := (1, \ldots, 1)^T$.

For every $x, y \in \Delta_n$ such that $x > 0$, have

$$\begin{aligned} dw_x(y) &= \sum_{i=1}^{n} [y_i \log y_i - x_i \log x_i - (1 + \log x_i)(y_i - x_i)] \\ &= \sum_{i=1}^{n} [y_i \log y_i - x_i \log x_i + (y_i - x_i) - (y_i - x_i) \log x_i] \\ &= \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{x_i} \right) + (y_i - x_i) \right] = \sum_{i=1}^{n} y_i \log \frac{y_i}{x_i} \end{aligned}$$

**Proposition 1.4** *Assume that* $\psi \in \overline{\text{Conv}}\,(\mathbb{R}^n)$ *and* $w$ *is a* $\nu$*-distance generating function for* $\operatorname{dom}\psi$*. Then,*

$$\inf\{(\psi + w)(x) : x \in \mathbb{R}^n\} \tag{1}$$

*has a unique optimal solution* $\bar{x}$*. Moreover, it holds that*

$$\bar{x} \in \operatorname{dom}\psi \cap W^0$$

**Proof**: Since $\psi, w \in \overline{\text{Conv}}\,(\mathbb{R}^n)$ and $\operatorname{dom}\psi \cap \operatorname{dom} w \neq \emptyset$, it follows that $\psi + w \in \overline{\text{Conv}}\,(\mathbb{R}^n)$. Moreover, since $w$ is $\nu$-strongly convex, it follows that $\psi + w$ is also $\nu$-strongly convex. Hence, (1) has a unique optimal solution $\bar{x}$. Clearly, $\bar{x} \in \operatorname{dom}\psi$. The optimality condition for (1) implies that

$$0 \in \partial(\psi + w)(\bar{x}) = \partial\psi(\bar{x}) + \partial w(\bar{x})$$

where the last equality is due to the fact that

$$\operatorname{ri}(\operatorname{dom}\psi) \cap \operatorname{ri}(\operatorname{dom} w) = \operatorname{ri}(\operatorname{dom}\psi) \cap W^0 = \operatorname{ri}(\operatorname{dom}\psi) \neq \emptyset$$

The above conclusion implies that $\partial w(\bar{x}) \neq \emptyset$, and hence that $\bar{x} \in W^0$ due to Definition 1.1(i).

## 2 Problem, assumptions and algorithm

Consider the optimization problem

$$\phi_* = \min\{\phi(x) := (f + h)(x) : x \in \mathbb{R}^n\} \tag{2}$$

where the following assumptions hold:

- $h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$
- $f \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ is such that $\mathrm{dom}\, f \supset \mathrm{dom}\, h$
- there exists a function $s : \mathrm{dom}\, h \to \mathbb{R}^n$ satisfying the following properties:
  - $s(x) \in \partial f(x)$ for all $x \in \mathrm{dom}\, h$
  - there exists $M \geq 0$ such that for every $x \in \mathrm{dom}\, h$,

$$\|s(x)\|_* \leq M \tag{3}$$

- optimal solution set $X_*$ is nonempty, and hence $\phi_* \in \mathbb{R}$

The second assumption above implies that

$$|f(x') - f(x)| \leq M\|x' - x\| \quad \forall x, x' \in \mathrm{dom}\, h.$$

Assume that $w$ is a $\nu$-distance generating function for $\operatorname{dom} h$. Observe that the definition of such function implies that

$$\operatorname{ri}(\operatorname{dom} h) \subset W^0, \quad \operatorname{dom} h \subset \operatorname{dom} w$$

where $W^0 := \operatorname{int}(\operatorname{dom} w)$.

---

**Mirror Descent Method (MDM)**

---

0) Let $x_0 \in W^0 \cap \operatorname{dom} h$ be given
1) For $k = 1, 2, \ldots$, do
   - set $s_{k-1} = s(x_{k-1})$
   - choose $\lambda_k > 0$ and let $x_k$ be the optimal solution of

$$\min \left\{ \ell_f(u; x_{k-1}) + h(u) + \frac{1}{\lambda_k} dw_{x_{k-1}}(u) \right\} \tag{4}$$

   where

$$\ell_f(\cdot; x_{k-1}) = f(x_{k-1}) + \langle s_{k-1}, \cdot - x_{k-1} \rangle$$

---

**Remark:** The objective function of (4) is well-defined as long as $x_{k-1} \in W^0 \cap \operatorname{dom} h$.

**Proposition 2.1** *If $x_{k-1} \in W^0 \cap \operatorname{dom} h$ then $x_k \in W^0 \cap \operatorname{dom} h$. Thus, MDM is well-defined.*

**Proof**: Follows from Proposition 1.4 with

$$\psi(\cdot) = \lambda_k [\ell_f(\cdot; x_{k-1}) + h(\cdot)] - \ell_w(\cdot; x_{k-1})$$

and the facts that $\operatorname{dom} \psi = \operatorname{dom} h$ and

$$x_k = \operatorname{argmin} \{ \psi + w)(x) : x \in \mathbb{R}^n \}$$

**Lemma 2.2** *For every $k \geq 1$,*

$$\frac{\nabla w(x_{k-1}) - \nabla w(x_k)}{\lambda_k} \in s_{k-1} + \partial h(x_k)$$

**Proof**: The optimality condition for (4) implies that

$$\begin{aligned}
0 &\in \partial \left( \ell_f(\cdot; x_{k-1}) + h(\cdot) + \frac{1}{\lambda_k} dw_{x_{k-1}}(\cdot) \right) (x_k) \\
&= s_{k-1} - \frac{1}{\lambda_k} \nabla w(x_{k-1}) + \partial \left( h(\cdot) + \frac{1}{\lambda_k} w(\cdot) \right) (x_k) \\
&= s_{k-1} - \frac{1}{\lambda_k} [\nabla w(x_{k-1}) - \nabla w(x_k)] + \partial h(x_k)
\end{aligned}$$

where the last equality is due to the fact that

$$\mathrm{ri}\,(\mathrm{dom}\, w) \cap \mathrm{ri}\,(\mathrm{dom}\, h) = W^0 \cap \mathrm{ri}\,(\mathrm{dom}\, h) = \mathrm{ri}\,(\mathrm{dom}\, h) \neq \emptyset$$

**Lemma 2.3** *For every $k \geq 1$ and $u \in \operatorname{dom} w$,*

$$
\begin{aligned}
dw_{x_{k-1}}(u) - dw_{x_k}(u) \geq\ & dw_{x_{k-1}}(x_k) - \lambda_k M \|x_k - x_{k-1}\| \\
& + \lambda_k \langle s_{k-1}, x_{k-1} - u \rangle + h(x_k) - h(u)
\end{aligned}
$$

**Proof**: To simplify notation, let $z_0 = x_{k-1}$, $z = x_k$, $s_f^0 = s_{k-1}$, and $\lambda = \lambda_k$. By Lemma 2.2, have

$$
s_h := \frac{\nabla w(z_0) - \nabla w(z)}{\lambda} - s_f^0 \in \partial h(z)
$$

Have

$$
\begin{aligned}
& dw_{x_{k-1}}(u) - dw_{x_k}(u) \\
& \qquad = dw_{z_0}(u) - dw_z(u) \\
\text{(Lemma 1.2)} \quad & \qquad = dw_{z_0}(z) + \langle \nabla w(z) - \nabla w(z_0), u - z \rangle \\
& \qquad = dw_{z_0}(z) + \langle \nabla w(z_0) - \nabla w(z), z - u \rangle \\
\text{(def of } s_h) \quad & \qquad = dw_{z_0}(z) + \langle \lambda(s_f^0 + s_h), z - u \rangle \\
& \qquad = dw_{z_0}(z) + \langle \lambda s_f^0, z - u \rangle + \langle \lambda s_h, z - u \rangle \\
& \qquad = \left[ dw_{z_0}(z) + \langle \lambda s_f^0, z - z_0 \rangle \right] + \langle \lambda s_f^0, z_0 - u \rangle + \langle \lambda s_h, z - u \rangle \\
& \qquad = \left[ dw_{z_0}(z) + \langle \lambda s_f^0, z - z_0 \rangle \right] + \langle \lambda s_f^0, z_0 - u \rangle + \lambda [h(z) - h(u)] \\
& \qquad \geq \left[ dw_{z_0}(z) - \lambda \|s_f^0\|_* \, \|z - z_0\| \right] + \langle \lambda s_f^0, z_0 - u \rangle + \lambda [h(z) - h(u)] \\
& \qquad \geq [dw_{z_0}(z) - \lambda M \|z - z_0\|] + \langle \lambda s_f^0, z_0 - u \rangle + \lambda [h(z) - h(u)]
\end{aligned}
$$

**Lemma 2.4** *For every $k \geq 1$ and $u \in \operatorname{dom} w$,*

$$2\lambda_k^2 \nu M^2 + dw_{x_{k-1}}(u) - dw_{x_k}(u) \geq \lambda_k[\phi(x_k) - \phi(u)]$$

**Proof**: For every $k \geq 1$ and $u \in \operatorname{dom} w$, have

$$\begin{aligned}
dw_{x_{k-1}}(u) - dw_{x_k}(u) &= dw_{z_0}(u) - dw_z(u) \\
&\geq [dw_{z_0}(z) - \lambda M\|z - z_0\|] + \langle \lambda s_f^0, z_0 - u\rangle \\
&\quad + \lambda[h(z) - h(u)] \\
&\geq [dw_{z_0}(z) - \lambda M\|z - z_0\|] + \lambda[f(z_0) - f(u)] \\
&\quad + \lambda[h(z) - h(u)] \\
&= [dw_{z_0}(z) - \lambda M\|z - z_0\|] + \lambda[f(z_0) - f(z)] \\
&\quad + \lambda[(f+h)(z) - (f+h)(u)] \\
&\geq [dw_{z_0}(z) - \lambda M\|z - z_0\|] - \lambda M\|z - z_0\| \\
&\quad + \lambda[\phi(z) - \phi(u)] \\
&\geq [dw_{z_0}(z) - 2\lambda M\|z - z_0\|] + \lambda[\phi(z) - \phi(u)] \\
&\geq \left[\frac{\nu\|z - z_0\|^2}{2} - 2\lambda M\|z - z_0\|\right] + \lambda[\phi(z) - \phi(u)] \\
&\geq -2\lambda^2\nu M^2 + \lambda[\phi(z) - \phi(u)]
\end{aligned}$$

**Lemma 2.5** *For every $K \geq 1$, $u \in \operatorname{dom} w$, and point $\bar{x}_K$ such that*

$$\phi(\bar{x}_K) \leq \frac{\sum_{k=1}^K \lambda_k \phi(x_k)}{\Lambda_K},$$

*we have*

$$\phi(\bar{x}_K) - \phi(u) \leq \frac{2M^2\nu \sum_{k=1}^K \lambda_k^2 + [dw_{x_0}(u) - dw_{x_K}(u)]}{\Lambda_K}$$

**Proof**: It follows from Lemma 2.4 that

$$\sum_{k=1}^K \lambda_k[\phi(x_k) - \phi(u)] \leq 2M^2\nu \sum_{k=1}^K \lambda_k^2 + [dw_{x_0}(u) - dw_{x_K}(u)]$$

This together with the assumption on $\bar{x}_K$ and the definition of $\Lambda_K$ imply the result.

**Proposition 2.6** *For every $K \geq 1$,*

$$\phi(\bar{x}_K) - \phi_* \leq \frac{2M^2\nu \sum_{k=1}^{K} \lambda_k^2 + dw_{x_0}(x_*)}{\Lambda_K}$$

$$dw_{x_K}(x_*) \leq dw_{x_0}(x_*) + 2M^2\nu \sum_{k=1}^{K} \lambda_k^2$$

**Proof**: Follows from Lemma 2.5 with $u = x_*$.

**Proposition 2.7 (Constant stepsize)** *Assume that*

$$\lambda_k = \lambda = \frac{\varepsilon}{4\nu M^2} \quad \forall k \geq 1$$

*Then, for any*

$$K \geq \frac{\nu M^2 D_0}{8\varepsilon^2} \tag{5}$$

*where $D_0 := \inf\{dw_{x_0}(x_*) : x_* \in X_*\}$, we have*

$$\phi(\bar{x}_K) - \phi_* \leq \varepsilon$$

**Proof**: For any $K$ satisfying (5), have

$$\begin{aligned}\phi(\bar{x}_K) - \phi_* &\leq \frac{2M^2\nu \sum_{k=1}^{K} \lambda_k^2 + D_0}{\Lambda_K} = \frac{2M^2\nu K\lambda^2 + D_0}{K\lambda} = 2M^2\nu\lambda + \frac{D_0}{K\lambda} \\ &= \frac{\varepsilon}{2} + \frac{4\nu M^2 D_0}{K\varepsilon} \leq \varepsilon\end{aligned}$$

■

Hence, the $\varepsilon$-iteration-complexity of MDM is

$$\mathcal{O}\left(\frac{\nu M^2 D_0}{\varepsilon^2}\right)$$

## 3 Application

Consider the optimization problem (2) where $h(\cdot)$ is the indicator of

$$X = \Delta_n := \{x \in \mathbb{R}^n_+ : \langle e, x\rangle = 1\}$$

where $e = (1, \ldots, 1)^T$. Take $x_0 = e/n$.

**Euclidean setting:** Choose

$$w(x) = \frac{1}{2}\langle x, x\rangle, \quad \|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}.$$

Then

$$\|\cdot\|_* = \|\cdot\|, \quad \nu = 1$$

For any $x \in \Delta_n$, have

$$\begin{aligned} dw_{x_0}(x) = \frac{1}{2}\|x - x_0\|^2 = \frac{1}{2}\|x - (e/n)\|^2 &= \frac{1}{2}\left(\|x\|^2 - \frac{2}{n}\langle e, x\rangle + \frac{1}{n^2}\|e\|^2\right) \\ = \frac{1}{2}\left(\|x\|^2 - \frac{2}{n} + \frac{1}{n}\right) &= \frac{1}{2}\left(\|x\|^2 - \frac{1}{n}\right) \le \frac{1}{2}\left(1 - \frac{1}{n}\right) \le \frac{1}{2} \end{aligned}$$

The Euclidean version of MDM has $\varepsilon$-iteration-complexity equal to

$$\mathcal{O}\left(\frac{M_2^2}{\varepsilon^2}\right)$$

where

$$M_2 = \sup\{\|s(x)\|_2 : x \in \Delta_n\}$$

and $\|\cdot\|_2$ is usual Euclidean norm.

**Non-Euclidean setting:** Choose

$$w(x) = \frac{1}{2}\sum_{i=1}^{n} x_i \log x_i, \quad \|x\| := \|x\|_1 := \sum_{i=1}^{n} |x_i|$$

Then

$$\|\cdot\|_* = \|\cdot\|_\infty, \quad \nu = 1$$

For every $x, y \in \Delta_n$ such that $x > 0$, have

$$dw_x(y) = \sum_{i=1}^{n} y_i \log \frac{y_i}{x_i}$$

Hence, for any $u \in \Delta_n$, have

$$dw_{x_0}(u) = \sum_{i=1}^{n} u_i \log(nu_i) = \log n + \sum_{i=1} u_i \log u_i \leq \log n$$

Recall that $w$ is 1-strongly convex on $\Delta_n$ with respect to $\|\cdot\|_1$

MDM has $\varepsilon$-iteration-complexity equal to

$$\mathcal{O}\left(\frac{M_\infty^2 \log n}{\varepsilon^2}\right)$$

where

$$M_\infty = \sup\{\|s(x)\|_\infty : x \in \Delta_n\}$$

**Comparison**: The ratio between the two complexities is

$$R := \left(\frac{M_\infty}{M_2}\right)^2 \log n$$

which satisfies

$$\frac{\log n}{n} \leq R \leq \log n$$

In practice, $R$ is closer to the lower bound than it is to upper bound, which generally favors the non-Euclidean version of MDM.

**Remark:** The solution of the prox subproblem (4) in the non-Euclidean version of MDM has a closed form, namely,

$$(x_k)_i = \frac{(x_{k-1})_i \exp[-\lambda_k (s_{k-1})_i]}{\sum_{i=1}^n (x_{k-1})_i \exp[-\lambda_k (s_{k-1})_i]} > 0$$

while in the Euclidean setting a (usually inexpensive) line search needs to be performed to compute $x_k$.