Mirror Descent

Axel Böhm

November 4, 2021

Recap on (sub)-gradient descent

- When we used a norm $\|\cdot\|$ we meant the 2-*norm*, i.e.

$$\|x\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}.$$

- In **gradient descent** we used Lipschitz continuity:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

(Lead to a complexity of $\mathcal{O}(\frac{L}{k})$)

- For **sub-gradient descent** we used $\|g\| \leq G$ which lead to a complexity of $\mathcal{O}(\frac{G}{\sqrt{k}})$.
- But there are other norms

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$$

It can happen that $\|g\|_\infty \leq G$ but $\|g\|_2 \approx \sqrt{d}G$.

Recap on (sub)-gradient descent

- When we used a norm $\|\cdot\|$ we meant the 2-*norm*, i.e.

$$\|x\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}.$$

- In **gradient descent** we used Lipschitz continuity:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

(Lead to a complexity of $\mathcal{O}(\frac{L}{k})$)

- For **sub-gradient descent** we used $\|g\| \leq G$ which lead to a complexity of $\mathcal{O}(\frac{G}{\sqrt{k}})$.
- But there are other norms

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$$

It can happen that $\|g\|_\infty \leq G$ but $\|g\|_2 \approx \sqrt{d}G$.

Recap on (sub)-gradient descent

- When we used a norm $\|\cdot\|$ we meant the 2-*norm*, i.e.

$$\|x\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}.$$

- In **gradient descent** we used Lipschitz continuity:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

(Lead to a complexity of $\mathcal{O}(\frac{L}{k})$)

- For **sub-gradient descent** we used $\|g\| \leq G$ which lead to a complexity of $\mathcal{O}(\frac{G}{\sqrt{k}})$.
- But there are other norms

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$$

It can happen that $\|g\|_\infty \leq G$ but $\|g\|_2 \approx \sqrt{d}G$.

Different norms?

But where did we use the norm in the ***method****?*

Gradient Descent

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

equivalently

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}$$

We can replace the 2-norm with a more general **distance**.

Bregman distance

$h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex

(i) h is differentiable of the interior of dom h

(ii) h is 1-strongly convex w.r.t. $\|\cdot\|_2$

Then

$$\mathcal{D}_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

Properties

- $\mathcal{D}_h(x, y) \geq 0$
- $\mathcal{D}_h(x, y) \neq \mathcal{D}_h(y, x)$
- $\mathcal{D}_h(\cdot, y)$ is convex for all y

$$\mathcal{D}_h(x, y) \approx \frac{1}{2}\langle \nabla^2 h(y)(x - y), x - y \rangle = \frac{1}{2}\|x - y\|^2_{\nabla^2 h(y)}$$

- $\mathcal{D}_h(x, y) \geq \frac{1}{2}\|x - y\|^2$ (1-strong convexity)

Bregman distance

$h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex

(i) h is differentiable of the interior of dom h

(ii) h is 1-strongly convex w.r.t. $\|\cdot\|_2$

Then

$$\mathcal{D}_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

Properties

- $\mathcal{D}_h(x, y) \geq 0$
- $\mathcal{D}_h(x, y) \neq \mathcal{D}_h(y, x)$
- $\mathcal{D}_h(\cdot, y)$ is convex for all y

$$\mathcal{D}_h(x, y) \approx \frac{1}{2}\langle \nabla^2 h(y)(x - y), x - y \rangle = \frac{1}{2}\|x - y\|^2_{\nabla^2 h(y)}$$

- $\mathcal{D}_h(x, y) \geq \frac{1}{2}\|x - y\|^2$ (1-strong convexity)

Examples

- $h(x) = \frac{1}{2}\|x\|_2^2$ gives $\mathcal{D}_h(x, y) = \|x - y\|^2$
- $h(x) = \frac{1}{2(p-1)}\|x\|_p^2$ with $p \in [1, 2]$
- $\Delta^d = \{x \in \mathbb{R}_+^d : \sum_{i=1}^d x_i = 1\}$ the *unit simplex* and

$$h(x) = \begin{cases} \sum_{i=1}^d x_i \log(x_i) & x_i > 0 \\ +\infty & \text{otherwise} \end{cases}$$

the **Negative entropy**.

Negative entropy

- Negative entropy: $h(x) = \sum_{i=1}^{d} x_i \log(x_i)$ for $x_i > 0$.
- Then $\nabla h(x) = \log(x) + 1$ (coordinatewise) and

$$\begin{aligned}
\mathcal{D}_h(x, y) &= \sum_{i=1}^{d} x_i \log(x_i) - y_i \log(y_i) - \langle \log(y) + 1, x - y \rangle \\
&= \sum_{i=1}^{d} x_i \log(x_i) - \sum_{i=1}^{d} x_i \log(y_i) \\
&= \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{y_i}\right)
\end{aligned}$$

  Known as Kullback-Leibler divergence $K(X \| Y)$.
- Is strongly convex over Δ

$$\mathcal{D}(x, y) \geq \frac{1}{2} \|x - y\|_1^2 \quad \text{Pinsker's ineq.}$$

Mirror descent

Idea: replace squared Euclidian norm with more general object:

$$
\begin{aligned}
x_{k+1} &= \arg\min_{x\in\mathbb{R}^d}\{f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\alpha_k}\mathcal{D}_h(x, x_k)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}\mathcal{D}_h(x, x_k)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}(h(x) - h(x_k) - \langle \nabla h(x_k), x - x_k\rangle)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}(h(x) - \langle \nabla h(x_k), x\rangle)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \alpha_k \nabla f(x_k) - \nabla h(x_k), x\rangle + h(x)\}
\end{aligned}
$$

Question: But why *mirror* descent?

Mirror descent

Idea: replace squared Euclidian norm with more general object:

$$
\begin{aligned}
x_{k+1} &= \arg\min_{x\in\mathbb{R}^d}\{f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\alpha_k}\mathcal{D}_h(x, x_k)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}\mathcal{D}_h(x, x_k)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}(h(x) - h(x_k) - \langle \nabla h(x_k), x - x_k\rangle)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \nabla f(x_k), x\rangle + \frac{1}{\alpha_k}(h(x) - \langle \nabla h(x_k), x\rangle)\} \\
&= \arg\min_{x\in\mathbb{R}^d}\{\langle \alpha_k \nabla f(x_k) - \nabla h(x_k), x\rangle + h(x)\}
\end{aligned}
$$

Question: But why *mirror* descent?

Mirror descent

Idea: replace squared Euclidian norm with more general object:

$$
\begin{aligned}
x_{k+1} &= \arg\min_{x \in \mathbb{R}^d} \{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\alpha_k} \mathcal{D}_h(x, x_k) \} \\
&= \arg\min_{x \in \mathbb{R}^d} \{ \langle \nabla f(x_k), x \rangle + \frac{1}{\alpha_k} \mathcal{D}_h(x, x_k) \} \\
&= \arg\min_{x \in \mathbb{R}^d} \{ \langle \nabla f(x_k), x \rangle + \frac{1}{\alpha_k} (h(x) - h(x_k) - \langle \nabla h(x_k), x - x_k \rangle) \} \\
&= \arg\min_{x \in \mathbb{R}^d} \{ \langle \nabla f(x_k), x \rangle + \frac{1}{\alpha_k} (h(x) - \langle \nabla h(x_k), x \rangle) \} \\
&= \arg\min_{x \in \mathbb{R}^d} \{ \langle \alpha_k \nabla f(x_k) - \nabla h(x_k), x \rangle + h(x) \}
\end{aligned}
$$

Question: But why *mirror* descent?

The Mirror part

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \{\langle \alpha_k \nabla f(x_k) - \nabla h(x_k), x \rangle + h(x)\}$$

By optimality condition:

$$0 = \alpha_k \nabla f(x_k) - \nabla h(x_k) + \nabla h(x_{k+1})$$

Therefore

$$\nabla h(x_{k+1}) = \nabla h(x_k) - \alpha_k \nabla f(x_k)$$

Why it's called mirror descent

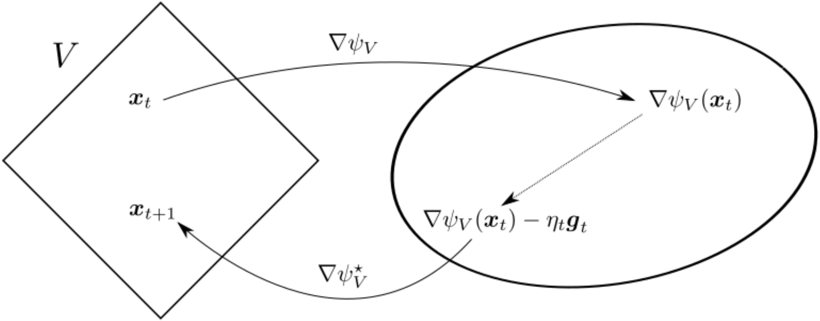

Figure: $\psi = h$

Mirror Descent on the unit simplex

Negative entropy: $h(x) = \sum_{i=1}^{d} x_i \log(x_i)$ for $x_i > 0$.
We define $a := \alpha_k \nabla f(x_k) - \nabla h(x_k)$. Then

$$x_{k+1} = \arg\min_{x \in \Delta} \{\langle a, x \rangle + h(x)\}$$

with $x_i \geq 0$ and $\sum x_i = 1$.

How to solve this?

Via **Lagrange**

$$L(x, \mu) = \langle a, x \rangle + h(x) - \mu(x_1 + \cdots + x_d - 1)$$

Mirror Descent on the unit simplex [contd]

Then,

$$\begin{aligned} \partial_{x_i} L(x,\mu) &= a_i + \log(x_i) + 1 - \mu \overset{!}{=} 0 \\ \log(x_i) &= \mu - 1 - a_i \\ x_i &= e^{\mu - 1 - a_i} = \beta e^{-a_i} \end{aligned}$$

with $\beta = e^{\mu-1}$.

Second constraint:

$$\sum_{i=1}^{d} x_i \overset{!}{=} 1 \Rightarrow \sum_{i=1}^{d} \beta e^{-a_i} = 1 \Rightarrow \beta = \frac{1}{\sum_{i=1}^{d} e^{-a_i}} \Rightarrow x_i = \frac{e^{-a_i}}{\sum_{j=1}^{d} e^{-a_j}}$$

Final mirror descent update:

$$x_{k+1}(i) = \frac{x_k(i) e^{\alpha_k [\nabla f(x_k)]_i}}{\sum_{j=1}^{d} e^{\alpha_k [\nabla f(x_k)]_j}}$$

Mirror Descent on the unit simplex [contd]

Then,

$$\begin{aligned}\partial_{x_i} L(x,\mu) &= a_i + \log(x_i) + 1 - \mu \overset{!}{=} 0 \\ \log(x_i) &= \mu - 1 - a_i \\ x_i &= e^{\mu - 1 - a_i} = \beta e^{-a_i}\end{aligned}$$

with $\beta = e^{\mu-1}$.
Second constraint:

$$\sum_{i=1}^{d} x_i \overset{!}{=} 1 \Rightarrow \sum_{i=1}^{d} \beta e^{-a_i} = 1 \Rightarrow \beta = \frac{1}{\sum_{i=1}^{d} e^{-a_i}} \Rightarrow x_i = \frac{e^{-a_i}}{\sum_{j=1}^{d} e^{-a_j}}$$

Final mirror descent update:

$$x_{k+1}(i) = \frac{x_k(i) e^{\alpha_k [\nabla f(x_k)]_i}}{\sum_{j=1}^{d} e^{\alpha_k [\nabla f(x_k)]_j}}$$

Mirror Descent on the unit simplex [contd]

Then,

$$\begin{aligned}
\partial_{x_i} L(x,\mu) &= a_i + \log(x_i) + 1 - \mu \overset{!}{=} 0 \\
\log(x_i) &= \mu - 1 - a_i \\
x_i &= e^{\mu - 1 - a_i} = \beta e^{-a_i}
\end{aligned}$$

with $\beta = e^{\mu-1}$.
Second constraint:

$$\sum_{i=1}^{d} x_i \overset{!}{=} 1 \Rightarrow \sum_{i=1}^{d} \beta e^{-a_i} = 1 \Rightarrow \beta = \frac{1}{\sum_{i=1}^{d} e^{-a_i}} \Rightarrow x_i = \frac{e^{-a_i}}{\sum_{j=1}^{d} e^{-a_j}}$$

Final mirror descent update:

$$x_{k+1}(i) = \frac{x_k(i) e^{\alpha_k [\nabla f(x_k)]_i}}{\sum_{j=1}^{d} e^{\alpha_k [\nabla f(x_k)]_j}}$$

(General) mirror descent convergence statement

Since we changed norm in the space of the variable x, we need to go to the dual norms in the space of the subgradients

$$\|y\|_* := \max_{\|x\|=1} \{\langle y, x \rangle\}.$$

Theorem

In $(\mathbb{R}^d, \|\cdot\|)$ and subgradients bounded in dual norm $\|g_k\|_* \leq G$, then

$$f(\bar{x}_k) - f^* \leq \frac{(\mathcal{D}(x^*, x_0))^{1/2} G}{\sqrt{k}},$$

where $\bar{x}_k$ denotes the averaged iterates, as usual.

Convergence on the unit simplex

What about $\mathcal{D}(x^*, x_0)$? Let $x_0 = (\frac{1}{n}, \cdots, \frac{1}{n})$, then

$$\mathcal{D}(x, x_0) = \sum x_i \log\left(\frac{x_i}{\frac{1}{n}}\right) = \sum x_i \log(x_i) + \log(n) \leq log(n)$$

while $\|x_0 - x^*\|^2 \leq 2$.
But if

$$\|g\|_\infty = \|g\|_1^* \leq G$$

we can still have

$$\|g\|_2 \approx \sqrt{d}G.$$

Proof

In the Euclidian space we used

$$\langle x_{k+1} - x_k, x^* - x_{k+1}\rangle$$
$$= \frac{1}{2}\|x^* - x_k\|^2 - \frac{1}{2}\|x^* - x_{k+1}\|^2 - \frac{1}{2}\|x_{k+1} - x_k\|^2.$$

Similar 3-point identity holds for Bregman distances:

$$\langle \nabla h(x_{k+1}) - \nabla h(x_k), x^* - x_{k+1}\rangle =$$
$$= D(x^*, x_k) - D(x^*, x_{k+1}) - D(x_{k+1}, x_k).$$

Therefore

$$D(x^*, x_{k+1}) \leq D(x^*, x_k) - D(x_{k+1}, x_k) + \alpha\langle g_k, x^* - x_{k+1}\rangle.$$

Proof II

$$D(x^*, x_{k+1}) \leq D(x^*, x_k) - D(x_{k+1}, x_k) + \alpha\langle g_k, x^* - x_{k+1}\rangle.$$

Last term is not quite right.

$$\begin{aligned}\langle g_k, x^* - x_{k+1}\rangle &= \langle g_k, x^* - x_k\rangle + \langle g_k, x_k - x_{k+1}\rangle \\ &\leq f(x^*) - f(x_k) + \|g_k\|_*\|x_k - x_{k+1}\| \\ &\leq f(x^*) - f(x_k) + \frac{\alpha\|g_k\|_*^2}{2} + \frac{\|x_k - x_{k+1}\|^2}{2\alpha}.\end{aligned}$$

Combined we get that

$$\begin{aligned}D(x^*, x_{k+1}) \leq D(x^*, x_k) - D(x_{k+1}, x_k) + \alpha(f(x^*) - f(x_k)) \\ + \frac{\alpha^2\|g_k\|_*^2}{2} + \frac{\|x_k - x_{k+1}\|^2}{2}.\end{aligned}$$

Proof III

We assumed strong convexity of h:

$$D(x_{k+1}, x_k) \geq \frac{1}{2}\|x_{k+1} - x_k\|^2.$$

Yields

$$D(x^*, x_{k+1}) \leq D(x^*, x_k) + \alpha(f(x^*) - f(x_k)) + \frac{\alpha^2 \|g_k\|_*^2}{2}$$

Continue as always

$$\frac{1}{k}\sum_{i=1}^{k} f(x_i) - f^* \leq \frac{D(x^*, x_0)}{\alpha k} \frac{\alpha G^2}{2}$$

What about the smooth case

- Talked about how to get better constants in the "bounded subgradients" setting
- but can't make them bounded if they are not

However,

- Can also come up with a new notion of smoothness

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LD(y, x)$$

- which might hold even if f is not smooth in classical sense