Mirror Descent

Axel

September 6, 2021

Recap on (sub)-gradient descent

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

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

In *gradient descent* we used Lipschitz continuity:

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

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

Recap on (sub)-gradient descent

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

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

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

$$\|g\| \leq G$$

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

Recap on (sub)-gradient descent

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

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

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

$$\|g\| \leq G$$

(Lead to a complexity of $\mathcal{O}(\frac{G}{\sqrt{k}})$)
But there are other norms

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

But where did we use the norm in the method?

Gradient Descent

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

equivalently

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

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

Bregman distance

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

1. h is differentiable of the interior of dom h
2. h is 1-strongly convex w.r.t. $\|\cdot\|_2$

Then

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

Bregman distance

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

1. h is differentiable of the interior of dom h
2. 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

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

Examples

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

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

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

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

Mirror descent

Idea: replace squared Euclidian norm with more general object:

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

Mirror descent

Idea: replace squared Euclidian norm with more general object:

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

Mirror descent

Idea: replace squared Euclidian norm with more general object:

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

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

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

Put mirror picture here

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,

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

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

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:

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

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

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

Theorem

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

$$f(\hat{x}_k) - f^* \leq \frac{\mathcal{D}(x^*, x_1) + \frac{1}{2}\sum_i \alpha_i^2 G^2}{\sum_i \alpha_i}$$

(general) mirror descent convergence statement

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

Theorem

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

$$f(\hat{x}_k) - f^* \leq \frac{\mathcal{D}(x^*, x_1) + \frac{1}{2}\sum_i \alpha_i^2 G^2}{\sum_i \alpha_i}$$

What about $\mathcal{D}(x^*, x_1)$**?** Let $x_1 = (\frac{1}{n}, \cdots, \frac{1}{n})$, then

$$\mathcal{D}(x, x_1) = \sum x_i \log(\frac{x_i}{\frac{1}{n}}) = \sum x_i \log(x_i) + \log(n) \leq log(n)$$