Mirror Descent

Part 2: The algorithm

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : September 6, 2020
Last update : September 6, 2020

Bregman divergence

- A key ingredient of mirror descent is the Bregman divergence.
- Given a norm $\|\cdot\|$ (which can be any norm), the Bregman divergence defined on a function ϕ is defined as

$$B_h(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

  where ϕ is differentiable with respect to $\mathbf{x}$ and λ-strongly convex w.r.t. the norm $\|\cdot\|$:

$$\phi(\mathbf{x}) - \frac{\lambda}{2}\|\mathbf{x}\| \text{ is convex.}$$

- Based on the definition of strong-convexity,

$$B_\phi(\mathbf{x}, \mathbf{y}) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

- See here for more discussions on Bregman divergence, and see p.6 here for discussion on strongly convex function.

Some key notes on Bregman divergence

- Bregman divergence is not commutative: $B(\mathbf{x}, \mathbf{y}) \neq B(\mathbf{y}, \mathbf{x})$.
- Bregman divergence is nonnegative.
- Bregman divergence is like a generalization of the squared- Euclidean distance.
- See here for more discussions on Bregman divergence.

Bregman projection vs Euclidean projection

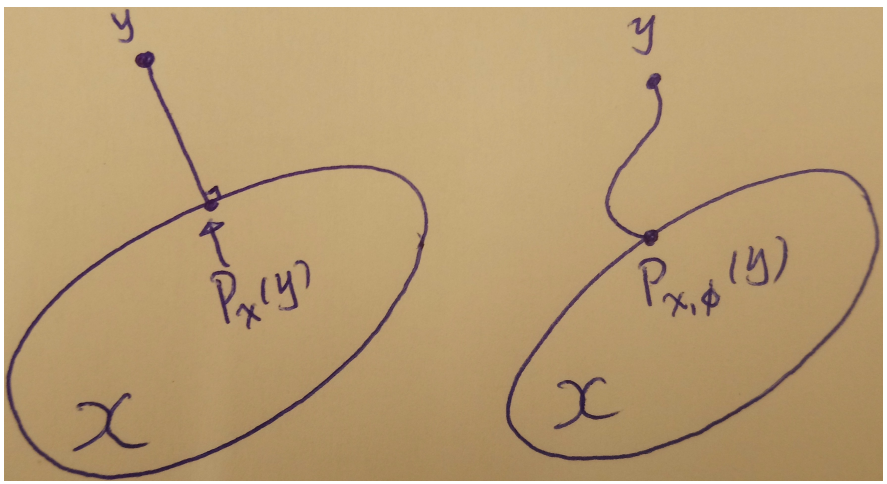
- Given a set $\mathcal{X}$, the Euclidean projection of a point $\mathbf{y}$ onto this set is defined as

$$P_{\mathcal{X}}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$$

- Given a set $\mathcal{X}$, the Bregman projection of a point $\mathbf{y}$ onto this set is, measured by the function ϕ, is defined as

$$P_{\mathcal{X},\phi}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} B_\phi(\mathbf{x},\mathbf{y}).$$

- Geometrically, Euclidean projection uses Euclidean geometry to give an orthogonal projection, while Bregman projection uses a curved geometry to give the projection.



Bregman projection vs Euclidean projection

- For Euclidean projection

$$\|\mathbf{x}-\mathbf{y}\|_2^2 \geq \|\mathbf{x}-P_{\mathcal{X}}(\mathbf{y})\|_2^2 + \|P_{\mathcal{X}}(\mathbf{y})-\mathbf{y}\|_2^2.$$

- Similarly, for Bregman projection,

$$B_\phi(\mathbf{x},\mathbf{y}) \geq B_\phi\Big(\mathbf{x}-P_{\mathcal{X}}(\mathbf{y})\Big) + B_\phi(P_{\mathcal{X}}(\mathbf{y})-\mathbf{y}).$$

- figure

The gradient descent algorithm

- The gradient descent update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{1}$$

can be treated as the minimization of a local quadratic model of f at the point $\mathbf{x}_k$

$$\begin{aligned} \mathbf{x}_{k+1} &= \operatorname*{argmin}_{\mathbf{x}} Q(\mathbf{x}, \mathbf{x}_k) \\ &= \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}_k\|_2^2. \end{aligned}$$

That is, if we take $\nabla Q(\mathbf{x}, \mathbf{x}_k) = 0$ gives (1).

- See here for more discussion.

The mirror descent algorithm

- In mirror descent, the L_2-norm term $\frac{1}{2}\|\cdot\|_2^2$ is replaced by the Bergman distance

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \; f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{\alpha} B_\phi(\mathbf{x}, \mathbf{x}_k).$$

  That's it, this update is the mirror descent !

- So what's the whole picture:
  1. You are given a function f to minimize
  2. You pick a norm $\|\cdot\|$ and a function ϕ that is λ-strongly convex w.r.t. the chosen $\|\cdot\|$.
  3. You minimize f using mirror descent.

  Note: you are free to choose any $\|\cdot\|$ and ϕ as you like, and when $\|\cdot\|$ and ϕ are well-chosen, mirror descent converges faster than gradient descent.

The mirror descent update

The mirror descent update

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{\alpha} B_\phi(\mathbf{x}, \mathbf{x}_k).$$

Remove constant term does not change the equation

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{\alpha} B_\phi(\mathbf{x}, \mathbf{x}_k).$$

Plug in the Bregman divergence

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{\phi(\mathbf{x}) - \phi(\mathbf{x}_k) - \langle \nabla \phi(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle.}{\alpha}$$

Remove constant terms again

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \langle \nabla f(\mathbf{x}_k), \mathbf{x} \rangle + \frac{\phi(\mathbf{x}) - \langle \nabla \phi(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle.}{\alpha}$$

Group similar term

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \langle \nabla f(\mathbf{x}_k) - \frac{1}{\alpha} \phi(\mathbf{x}_k), \mathbf{x} \rangle + \frac{1}{\alpha} \phi(\mathbf{x}).$$

We arrive at a compact expression of the mirror descent update.

The mirror descent algorithm

From the mirror descent update

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \underbrace{\langle \nabla f(\mathbf{x}_k) - \frac{1}{\alpha}\phi(\mathbf{x}_k), \mathbf{x}\rangle + \frac{1}{\alpha}\phi(\mathbf{x})}_{Q(\mathbf{x},\mathbf{x}_k)}, \qquad (*)$$

Similar to the gradient descent in slide 6, to find the solution of (*) , we take $\nabla Q(\mathbf{x}, \mathbf{x}_k) = 0$, this gives the update equation as

$$\nabla f(\mathbf{x}_k) - \frac{1}{\alpha}\phi(\mathbf{x}_k) + \frac{1}{\alpha}\nabla\phi(\mathbf{x}_{k+1}) = 0.$$

Rearrange

$$\nabla\phi(\mathbf{x}_{k+1}) = \phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k).$$

Viewing $\nabla\phi$ as an operator, we have

$$\mathbf{x}_{k+1} = \Big(\nabla\phi\Big)^{-1}\Big(\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)\Big).$$

As a side note, $\Big(\nabla\phi\Big)^{-1} = \nabla\Big(\phi^*\Big)$, where ϕ^* is the conjugate of ϕ.

Why mirror descent is called mirror descent

The mirror descent update

$$\mathbf{x}_{k+1} = \Big(\nabla\phi\Big)^{-1}\Big(\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)\Big).$$

$\nabla\phi$ and $(\nabla\phi)^{-1}$ are the mirror maps:

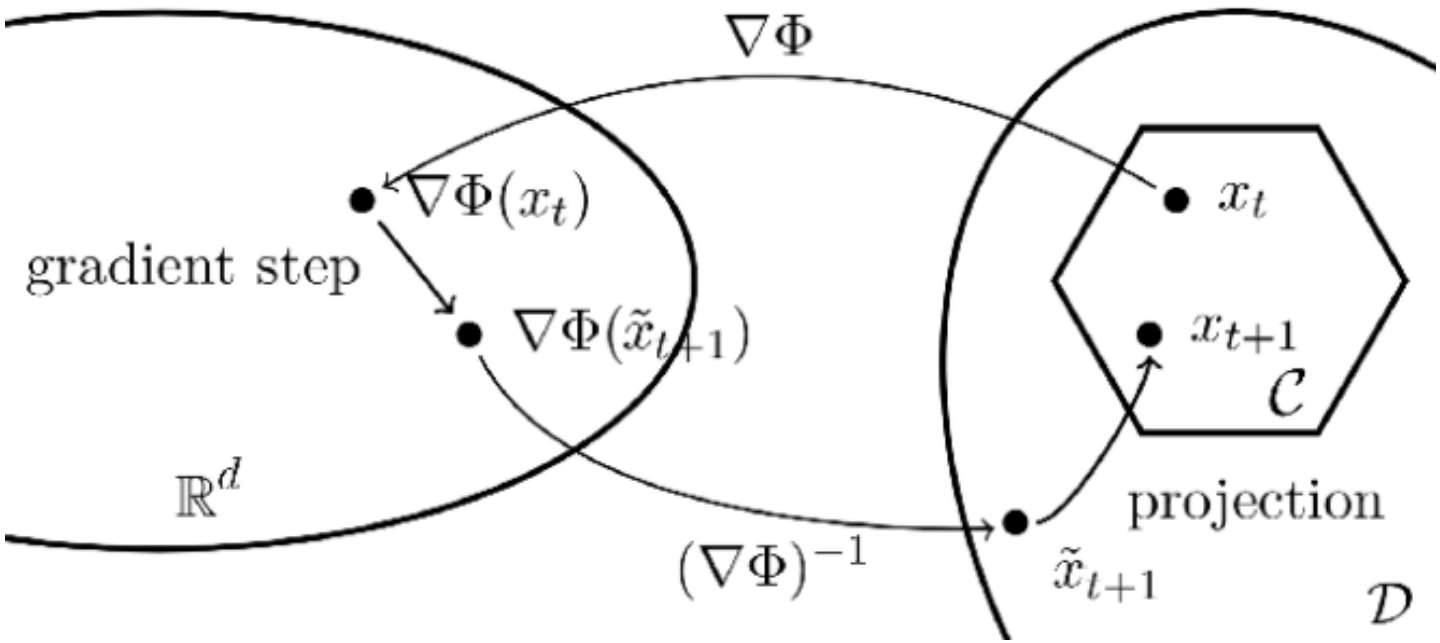


Figure: Geometry of mirror descent[1].

[1] Figure from S. Bubeck, Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning, 2015.

What about constrained problem

- The mirror descent update $\mathbf{x}_{k+1} = \Big(\nabla\phi\Big)^{-1}\Big(\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)\Big)$ we discussed is for unconstrained problem : because this update comes from replacing a term in the ordinary gradient descent step.

- For constrained problem $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$, the projected gradient descent adds a Euclidean projection on the gradient step

$$\mathbf{x}_{k+1} = P_{\mathcal{C}}\Big(\mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)\Big).$$

- for mirror descent, we can do the same: by adding a projection, we arrive at the projected mirror descent

$$\mathbf{x}_{k+1} = P_{\mathcal{C},\phi}\bigg(\Big(\nabla\phi\Big)^{-1}\Big(\phi(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)\Big).$$

  This corresponds to the $\tilde{\mathbf{x}}$ in the figure in the last slide.

- Important note: the projection in projected mirror descent is the Bregman projection, not Euclidean projection.

Last page - summary

- Bregman divergence
- Derivation of the mirror descent: replacing the L_2 norm in gradient descent by Bregman divergence.
- Different expression of the mirror descent update

End of document