Mirror Descent

Part 1: Why mirror descent? Why gradient descent is not enough?

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

A simplex-constrained problem

- Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, consider the minimization problem

$$(\mathcal{P}) \quad : \quad \min_{\mathbf{x} \in \Delta^n} f(\mathbf{x})$$

where Δ^n is the n-dimensional unit simplex:

$$\Delta^n := \left\{ \mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0, i = 1, 2, \ldots, m, \langle \mathbf{1}_n, \mathbf{x} \rangle = 1 \right\}.$$

- This document: using this problem as an example to explain the motivation of mirror descent.

Gradient-based method

$$(\mathcal{P}) \quad : \quad \min_{\mathbf{x} \in \Delta^n} f(\mathbf{x})$$

- How to solve this problem depends on the nature of f.
  - If f is differentiable: a natural way to solve $(\mathcal{P})$ is to use the projected gradient descent (ProjGD)

    $$\mathbf{x}_{k+1} = P_\Delta\Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\Big),$$

    where $\alpha_k > 0$ is the stepsize and P_Δ is the projection onto the simplex, see here for details on how to project onto the simplex.

  - If f is not differentiable: then we replace the notion of gradient by subgradient, and use the projected subgradient method (ProjSubGM)

    $$\mathbf{x}_{k+1} = P_\Delta\Big(\mathbf{x}_k - \alpha_k \mathbf{g}_k\Big),$$

    where $\mathbf{g}_k \in \partial f(\mathbf{x}_k)$ is a subgradient of f at the point $\mathbf{x}_k$, and $\partial f(\mathbf{x}_k)$ denotes the subdifferential of f at $\mathbf{x}_k$.

Convergence rate of projected subgradient method

- The convergence rate of the two gradient methods depend on the nature of f.
- If f is L-Lipschitz (i.e., the norm of subgradient or gradient is bounded above by a constant L), then the convergence rate is

$$f(\mathbf{x}_k) - f^* \leq c\frac{L}{\sqrt{k}}$$

  subject to some constant c.
- The above rate holds for projected gradient descent (see here) and also projected subgradient method.
- We are going to see that, mirror descent, while having the same denominator $\sqrt{k}$, has a better nominator, compared with ProjGD and ProjSubGM.

An additional condition

- Suppose $\|\partial f(\mathbf{x})\|_\infty \leq 1$.
  - This means, the largest element of any subgradient (thus including gradient) is bounded by 1.
  - This also means all the elements of any subgradient are bounded by 1.
  - The extreme cases are the following two vectors

$$\mathbf{1}_n = [1, 1, \ldots, 1]^\top, , \quad \mathbf{0}_n = [0, 0, \ldots, 0]^\top.$$

- For the vector $\mathbf{1}_n$, it has L_2-norm as $\sqrt{n}$. This means, the largest possible size (measured by L_2 norm) of any subgradient, under the condition $\|\partial f(\mathbf{x})\|_\infty \leq 1$, is $\sqrt{n}$, where n is the dimension of the vector.
- In other words, $\|\partial f(\mathbf{x})\|_\infty \leq 1$ gives $\|\partial f(\mathbf{x})\|_2 \leq L = \sqrt{n}$.
- Hence, if $\|\partial f(\mathbf{x})\|_\infty \leq 1$, solving $(\mathcal{P})$ by ProjGD or ProjSubGM, the converge rate is

$$f(\mathbf{x}_k) - f^* \leq c\sqrt{\frac{n}{k}}.$$

- It turns out the rate $\sqrt{\frac{n}{L}}$ is not optimal, mirror descent can do better as $\sqrt{\frac{\log n}{k}}$.

Last page - summary

- Why mirror descent
  - The $\sqrt{\frac{\log n}{k}}$ vs $\sqrt{\frac{n}{k}}$ example.
  - Mirror descent generalizes gradient descent from the Euclidean geometry to a "more general geometry": a Riemannian manifold. That is, by making use of the geometry (not just staying within the Euclidean geometry as always), the convergence can be improved.
- We will see: mirror descent update $=$ "gradient descent along a Riemannian manifold by multiplying the standard gradient by the inverse of the metric tensor".

End of document