# 关于Par VI的若干问题

WangFangyikang

2023 年 3 月 29 日

## 目录

# 1 简介

本文的目的在于使读者对ParVI采样方法形成初步了解，全文将从欧几里得空间上的梯度流入手，随后介绍概率空间上的梯度流。之后介绍ParVI方法，以及本研究组提出的DPVI框架。最后还会补充mirror DPVI方法的介绍。

预备知识：凸分析，测度论，泛函分析，多元微积分

本人学习整理若有疏忽，发现错误还望指正，请发送邮箱wangfangyikang@zju.edu.cn

# 2 在欧几里得空间上的梯度流

为了引出概率空间上的梯度流，本节首先从梯度下降视角引出欧几里得空间上的梯度流。随后将欧几里得空间上的思路推广至概率空间。考察一个经典的优化问题：

$$\underset{x\in\mathbb{R}^n}{min} f(x), f:\mathbb{R}^n \mapsto \mathbb{R}$$

在此问题上，有经典的梯度下降算法（Gradient Descent Algorithm）：

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

其中$\eta_k$为第$k$步迭代的步长，$\nabla f(x_k)$为$f(x)$在$x_k$这一点上的梯度（假设$f(x)$的性质足够好）。以上的迭代运算可被重写为：

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

上式可以看作如下ODE方程的显式欧拉插值算法每一步的迭代操作。

$$\frac{dX(t)}{dt} = -\nabla f(x_k) \tag{2}$$

由此，梯度下降算法可以看作是欧几里得空间上的梯度流的离散插值形式。现在先简单证明一下此梯度流可以保证，在$f:\mathbb{R}^n \to \mathbb{R}$为凸函数时，随机过程$X(t)$是趋近于最优点的。

证明. 令 $g(X(t) = \frac{1}{2}\|X(t) - x^*\|^2)$,其中$x^* = argminf(x)$,则有：

$$\begin{aligned}
\frac{dg(X(t))}{dt} &= \frac{dg(X(t))}{dX(t)}\frac{dX(t)}{dt} \\
&= \langle \frac{dX(t)}{dt}, X(t) - x^* \rangle \\
&= -\langle \nabla_x f(X(t)), X(t) - x^* \rangle \\
&= \langle \nabla_x f(X(t)), x^* - X(t) \rangle \\
&\leqslant f(x^*) - f(X(t)) \\
&\leqslant 0
\end{aligned} \tag{3}$$

第一个不等式来自于$f(x)$的凸性质。

可见$g(X(t))$随着时间$t$减小，$X(t)$逐步逼近$x^*$。 □

让我们再一次审视$\frac{dX(t)}{dt} = -\nabla f(x_k)$，这是一个处处由$-\nabla f(x)$ 定义的向量场。回顾多元微积分中梯度的概念，梯度存在的时候可以由 $\nabla f(x) = [\frac{\partial f(\cdot)}{\partial x_1}, \dots, \frac{\partial f(\cdot)}{\partial x_n}]$ 进行计算，但这并不是梯度这一概念的定义，更不是梯度这一概念的本质。接下来我将先介绍梯度的定义，而后介绍梯度的本质。只有理解了欧几里得空间中梯度的深层内涵，才有可能理解概率空间上的变分（variation）操作。

以$f:\mathbb{R}^n \mapsto \mathbb{R}$为例，首先定义方向导数

**定义 2.1.** 函数$f$在$\widehat{x}$这一点处关于方向$h \in \mathbb{R}^n$的方向导数定义为：

$$f'(\widehat{x}, h) = \lim_{t \to 0} \frac{f(\widehat{x} + th) - f(\widehat{x})}{t} \tag{4}$$

接下来给出加托可微性和梯度的定义。

**定义 2.2.** 当方向导数$f'(\widehat{x}, h), h \in \mathcal{X}$关于$h$是一个线性系统（满足数乘保持性和加法保持性）时，即 $f'(\widehat{x}, h) = \mathcal{A}(h)$（其中$\mathcal{A}$是某个线性算子）时，我们说$f$在$\widehat{x}$这一点加托-可微。

根据泛函分析中的里茨表示定理（Riesz representation theorem），$\mathbb{R}^n$ 空间上的线性算子构成的对偶空间与原始空间同构。考察$(\mathbb{R}^n, < \cdot, \cdot >_2)$这一Hilbert空间，则对于任意一个光滑的线性映射$l : \mathbb{R}^n \mapsto \mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$，有唯一一个$x_l \in \mathbb{R}^n$,使得$\forall x \in \mathbb{R}^n, l(x) = \langle x_l, x \rangle$。

所以$f : \mathbb{R}^n \mapsto \mathbb{R}$在某一点的梯度实际上是将加托-可微线性算子写作了一个向量，梯度的内涵是$f$在$\widehat{x}$邻域内的线性近似。回忆一下，这和我们当年学习的一阶泰勒展开是吻合的。

$$f(\widehat{x} + h) \approx f(x) + \mathcal{A}(h) + \mathcal{O}(\| h \|)$$

其中$\mathcal{A}$是一个光滑的线性算子，重写线性算子为内积形式，记作：

$$f(\widehat{x} + h) \approx f(x) + \langle \nabla f(\widehat{x}), h \rangle + \mathcal{O}(\| h \|)$$

从这一角度对梯度进行理解之后，我们可以模仿梯度的定义来定义概率分布的变分，进一步地，将梯度流拓展到概率空间，。