Bregman 散度

定义 1 (Bregman 散度)**.** 设 $f : \mathbb{R}^n \to \mathbb{R}$ 是一个凸函数，则对于任意两个点 $x, y \in \mathbb{R}^n$，*Bregman* 散度定义为：

$$D_f(x\|y) = f(x) - f(y) - \nabla f(y)^T(x - y)$$

其中，$\nabla f(y)$ 是函数 f 在点 y 的梯度。

理解

基本描述: Bregman 散度度量了点 x 相对于点 y 的 "远离" 程度。它通过考虑 f 在 y 处的线性近似，来量化从 y 到 x 的距离。由于 f 是凸的，$D_f(x\|y)$ 总是非负的，并且当且仅当 $x = y$ 时，$D_f(x\|y) = 0$。

Bregman 散度的几何解释与代数理解

几何上：Bregman 散度的大小意味着点 x 与参考点 y 之间的非线性差异，这种非线性差异也就是函数的凸性。如果 $D_f(x\|y)$ 较小，说明点 x 与点 y 在函数 f 下可以线性估计；而如果 $D_f(x\|y)$ 较大，则表明 x 与 y 之间存在显著的差异，也就是函数的凸性更强，函数的弯曲程度更高。

例如，图 1 中，蓝色曲线为线性函数，函数弯曲程度可视为 0，而红色曲线为二次函数，函数弯曲程度更高。

代数上：根据泰勒公式，对于一个在点 y 处可微的凸函数 f，我们有：

$$f(x) \approx f(y) + \nabla f(y)^T(x-y) + \frac{1}{2}(x-y)^T\nabla^2 f(y)(x-y),$$

其中 $\nabla^2 f(y)$ 是 Hessian 矩阵，反映了函数的二阶导数。Bregman 散度的差值与 Hessian 矩阵相关，Hessian 矩阵反映了函数梯度的变化速率。直观而言，Bregman 散度越大，意味着函数的梯度增长越快，函数的弯曲程度也越高。

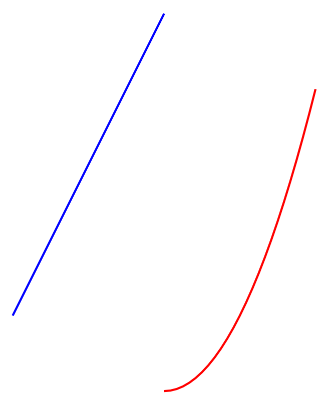

图 1: 线性函数与凸函数曲线

Bregman 散度与 KL 散度的关系

Bregman 散度和 Kullback-Leibler (KL) 散度之间存在密切的联系。实际上，KL 散度可以被视为 Bregman 散度的一种特殊情况。通过选择适当的凸函数，我们可以将 Bregman 散度转化为 KL 散度。

定理 1 (KL 散度作为 Bregman 散度). *设 $f(p) = -\sum_{i=1}^{k} p_i \log(p_i)$ 为一个适当的凸函数，其中 p_i 为离散概率分布的概率。则对应的 Bregman 散度 $D_f(p\|q)$ 可以表示为 KL 散度：*

$$D_f(p\|q) = D_{KL}(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

理解

在这个公式中，根据 Bregman 散度 $D_f(p\|q)$ 的解读为：当我们选择 Shannon 熵 f 作为凸函数时，Bregman 散度就变成了 KL 散度 $D_{\mathrm{KL}}(p \parallel q)$，因此衡量两个概率分布 p 和 q 之间的相似度，实际上是衡量 $f(p) = -\sum_{i=1}^{k} p_i \log(p_i)$ 的凸性，凸性越强，分布差异越大。
具体来说，KL 散度量化了从分布 q 到 p 的信息损失。它揭示了若模型使用 q 来逼近真实分布 p 时，所产生的额外的推理成本。通过这种连接，我们可以看到，Bregman 散度提供了一种更广泛的框架，使我们能够利用不同的凸函数进行距离测量，从而在优化和机器学习等领域发挥作用。

总结

Bregman 散度与 KL 散度之间的关系为我们提供了一个强有力的工具，可以在不同的上下文中应用。了解这两种散度的关系不仅有助于加深对凸优化的理解，还有助于在机器学习和统计学中设计更有效的算法。