非线性优化

一、状态估计问题

1.1 最大后验与最大似然

经典SLAM模型由一个状态方程和一个运动方程构成，如下式所示：

$$\begin{cases} x_k = f(x_{k-1}, u_k) + w_k \\ z_{k,j} = h(y_j, x_k) + v_{k,j} \end{cases}$$

由于运动方程在视觉SLAM中没有特殊性，目前不讨论，将观测方程具体化，由针孔相机成像原理得到：

$$sz_{k,j} = KTy_i$$

其中，s为像素点的距离（真实点到相机坐标系的距离Z），K为相机内参，T为相机位姿矩阵，yi为相机观测到的路标点在**世界坐标系**下的坐标，z为在xk处观测y时对应的图像像素位置。

考虑噪声影响，假设两个噪声项wk、vkj满足零均值的高斯(正态)分布。**在噪声的影响下，我们希望通过带噪声的数据z和u，推断位姿x和地图y（以及它们的概率分布）**，这构成了一个状态估计的问题。

状态变量：我们把所有待估计的变量放在一个“状态变量”中：

$$x = \{x_1, \dots, x_N, y_1, \dots, y_M\}$$

我们要解决的问题时是：已知输入数据（传感器读数）u和观测数据z（可以认为是像素位置）的条件下，计算状态x的**条件概率分布：**

$$P(x|z, u)$$

我们考虑没有测量运动的传感器时，只有图像，那么就计算P(x|z)，利用贝叶斯公式，有

$$P(x|z) = \frac{P(z|x)P(x)}{P(z)}$$

直接求后验概率P(x|z)较为困难，但是求一个状态最优估计，使得在该状态下，后验概率最大化（Maximize a Posterior，MAP），则是可行的。

$$x^*_{MAP} = argmaxP(x|z) = argmaxP(z|x)P(x)$$

对公式中arg的解释：arg max是使后面式子达到最大值时变量的取值。例如：函数F(x，y)，arg max是指当F(x，y)取得最大值时，变量x，y的取值

在不知道先验概率P（x）的情况下，可以求解x 的**最大似然估计**。即**在什么样的状态下，最可能产生现在观测到的数据。**

1.2 最小二乘法的引出

根据公式含义：

$$z_{k,j} = h(y_j, x_k) + v_{k,j}$$

上式表示，机器人在xk位置观测到路标yj，产生观测数据（像素）zkj，vkj为噪声，服从高斯分布。因此，我们可以得到观测数据的条件概率：

$$P(z_{j,k}|x_k,y_j)=N(h(y_j,x_k),Q_{k,j})$$

N代表服从正态分布，可以看出，这还是一个高斯分布。为了计算使它最大化的xk，yj（极大似然估计），**通常使用最小化负对数的方法。**

任意高维高斯分布：x~N（μ，∑）

$$P(x)=\frac{1}{\sqrt{(2\pi)^N det(\sum)}}e^{-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)}$$

其中，∑为相关系数，μ为均值。

取负对数后：前面系数与变量x无关，常数不管，只需要后面的二次型最小即可。可以使用无约束最优化求解（Newton、拟牛顿法、共轭梯度法、最速下降法）。

1.3 总结

首先，h（y，x）为定值，记为z0kj，但是由于噪声，观测值zkj与z0kj不一样。我们要求的是：在得到观测值z的情况下，x为和值的概率最大（例：z=(100,230)时，P（x=(1,2,3)^T）=0.95最大，那么认为x=(1,2,3)^T），即求P（x|z）。这个很不好求，因此使用贝叶斯公式转换，再极大似然估计，把x看作参数，观测位置为已知。现在变成求x为何值时，得到这个观测位置的概率最大，那么我们把这个x值认为位置，即求argmax(P（z|x）)，x为自变量。之后，这个最大似然估计的过程被转换为最小二乘问题。

二、非线性最小二乘法

最小二乘问题泛指具有如下形式的问题：

$$minf(x)=\frac{1}{2}\sum_{j=1}^{m}r_j^2(x)$$

其中，m一般指实例个数，rj指残差，即目标值和预估值的差。

我们想得到

2.1 一阶梯度与二阶梯度法

求解增量最直观的方式是将目标函数在x附近进行泰勒展开：

$$||f(x+\Delta x)||_2^2\approx||f(x)||_2^2+J(x)\Delta x+\frac{1}{2}\Delta x^T H\Delta x$$

我们可以看出，只保留一阶导数时(J雅可比矩阵)，增量方向显然与梯度方向相反：

$$\Delta x^*=-J^T(x)$$

而当保留二阶导数时，我们想得到增量Δx使得上式值最小，因此对Δx求导数并且使其为0，可以得到：

$$H\Delta x=-J^T$$

第一种方法称为最速下降法，第二种方法称为牛顿法

2.2 Gauss-Newton法(只能处理最小二乘问题)

整体思想：**将fx进行一阶泰勒展开：**

$$f(x+\Delta x)\approx f(x)+J(x)\Delta x$$

这里J(x)时fx关于x的导数，我们用一阶泰勒展开式代替原来的函数，再用得到的函数进行最小二乘求解。我们对一阶展开的上式求最小值：求Δx使得||f(x+Δx)||^2达到最小：

$$\Delta x^* = argmin_{\Delta x}\frac{1}{2}||f(x)+J(x)\Delta x||^2$$

我们对Δx求导(不是对x求导)，先化简得到：

$$\begin{aligned}&\frac{1}{2}||f(x)+J(x)\Delta x||^2\\ &=\frac{1}{2}(f(x)+J(x)\Delta x)^T(f(x)+J(x)\Delta x)\\ &=\frac{1}{2}(||f(x)||_2^2+2f(x)^TJ(x)\Delta x+\Delta x^TJ(x)^T\Delta x)\end{aligned}$$

对上式求导，并令导数为0：

$$2J(x)^Tf(x)+2J(x)^TJ(x)\Delta x=0$$

我们可以看到，上式提供了Δx的方程：

$$\begin{aligned}H(x)\Delta x = g(x)\\ \text{其中：}H=J(x)^TJ(x)\\ g=-J(x)^Tf(x)\end{aligned}$$

我们简化了牛顿法中计算Hesse矩阵的步骤，使用JTJ代替Hesse矩阵。因此，高斯牛顿法步骤如图：

1. 给定初始值 $\boldsymbol{x}_0$。
2. 对于第 k 次迭代，求出当前的雅可比矩阵 $\boldsymbol{J}(\boldsymbol{x}_k)$ 和误差 $f(\boldsymbol{x}_k)$。
3. 求解增量方程：$\boldsymbol{H}\Delta\boldsymbol{x}_k=\boldsymbol{g}$.
4. 若 $\Delta\boldsymbol{x}_k$ 足够小，则停止。否则，令 $\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\Delta\boldsymbol{x}_k$，返回 2.

2.3 Levenberg-Marquadt法（阻尼牛顿法）

根据泰勒公式，**只有当Δx足够小的时候，近似等式才有较好的效果。**因此，我们应该给Δx添加一个**信赖区域(Trust Region)**，不能让它太大而使得近似不准确。这种思想叫做：**信赖区域方法(Trust Region Method)**。在信赖区域中，我们认为近似有效；否则无效。

利用实际函数与近似函数的变化量的比值来确定信赖区域：

$$\rho=\frac{f(x+\Delta x)-f(x)}{J(x)\Delta x}$$

分子为实际函数的增量Δf，分子为一阶泰勒近似增量(可见2.2)。比值越接近于1，说明近似效果越好；比值太小，则应减小信赖区域范围；比值太大，则可以放大信赖区域范围。

1. 给定初始值 $\boldsymbol{x}_0$，以及初始优化半径 μ。
2. 对于第 k 次迭代，求解：

$$\min_{\Delta \boldsymbol{x}_k} \frac{1}{2}\|f(\boldsymbol{x}_k)+\boldsymbol{J}(\boldsymbol{x}_k)\Delta \boldsymbol{x}_k\|^2, \quad s.t. \|\boldsymbol{D}\Delta \boldsymbol{x}_k\|^2 \le \mu, \tag{6.24}$$

   这里 μ 是信赖区域的半径，$\boldsymbol{D}$ 将在后文说明。
3. 计算 ρ。
4. 若 $\rho > \frac{3}{4}$，则 $\mu = 2\mu$；
5. 若 $\rho < \frac{1}{4}$，则 $\mu = 0.5\mu$；
6. 如果 ρ 大于某阈值，认为近似可行。令 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta \boldsymbol{x}_k$。
7. 判断算法是否收敛。如不收敛则返回 2，否则结束。

上面是一个约束最优化问题，我们使用**Lagrange乘子**将其转换为一个无约束优化问题：

$$\Delta x^* = argmin_{\Delta x}\frac{1}{2}||f(x)+J(x)\Delta x||^2 + \frac{\lambda}{2}||D\Delta x||^2$$

类似高斯牛顿法展开，我们简单的得到：

$$(H+\lambda D^T D)\Delta x = g$$

可解方程。如果我们把D取为单位矩阵，那么有：

$$(H+\lambda I)\Delta x = g$$

当λ较大时，L-M法更接近于一阶梯度下降法；当λ较小时，L-M法更接近于高斯牛顿法。因此，LM法可以一定程度上避免线性方程组系数矩阵的奇异和病态问题。