高频交易的最优执行策略研究*

王　丹　向修海

内容提要:高频交易是最近金融行业非常热门的话题,如何在中国金融市场上实施高频交易更是备受大家关注。本文将介绍高频交易的一个非常重要组成部分——最优执行策略研究。本文首先剖析了最优执行策略的本质,即流动性或者流动性风险。然后基于此思想将最优执行策略模型划分为三代,并详细分析了三代模型之间的区别和联系。本文对最优执行策略模型发展历程的总结,既有利于激发未来的研究,又为中国高频交易的实施提供了实践启发。

关键词:高频交易　最优执行模型　流动性　流动性风险

一、引　言

高频交易是指投资者利用计算机自动捕捉市场上的每一个获利机会并生成程式化交易策略的交易方式。其最大特点是持仓时间短、交易量巨大。一般地,大额交易会对市场价格产生逆行冲击,扩大买卖价差,进而导致交易的平均执行价格异于市场可利用的最优价格。因此,高频交易的一个重要组成部分就是如何将大额交易分割为小额订单指令提交并执行,即最优执行策略研究。其实,最优执行策略不仅对高频交易很重要,而且在现实操作中,机构投资者面临同样的问题,如保险公司、证券自营商、基金公司等,他们的日常活动就是最优分割和执行其交易需求,具体地说,即分割后订单的执行时间和数量。正是实践的需要,最优执行理论的研究越来越热,已经形成了一个成熟完整的研究体系。但是目前国内外都还没有全面系统介绍该理论的文献综述,并且国内的相关研究仅仅围绕文中第一代模型而展开。有鉴于此,本文综述的目的之一是系统阐述该理论的历史发展沿革,以激发国内学者跟进这方面的研究,为推进中国的高频交易实践提供理论依据。

二、最优执行策略的本质:流动性或流动性风险

高频交易的变现或购买策略都是在短时间内完成预设的交易量。对它们来说,最优策略是将交易量分割成小额订单然后分批成交。然而从单笔交易来看,它们的交易会对当前均衡价格产生不利的影响,其均衡偏离幅度取决于当前最优价格附近的限价指令数量。这种偏离一般会产生三种效应,一是长期效应,也就是在整个交易周期内都存在,它主要通过中间价格的漂移而逐渐累积,对以后各个时刻的交易价格都产生影响;二是短暂效应,即它会在当前时刻的交易结束后立即衰减至零,因此,只对当前交易产生影响;三是跨期效应,它意味着均衡的偏离不会立即衰减至零,而是逐渐衰减至零。一般后二者总称为反弹效应。三种效应的综合会使得实际执行的平均价格不同于交易前可观测的价格,实际成交价格是交易规模、限价指令薄的形状、反弹效应的函数,而总执行成本是实际成交价与总交易计划的乘积。因此,最优执行策略问题就是求解使得总期望执行成本最小的交易策略。回顾Kyle(1985)对流动性的定义,这里实际成交价的影响因素与流动性的深度和反弹存在对应关系。因此,从流动性确定的角度来看,最优执行策略的本质正是流动性问题。如果我们进一步假设流动性是随机的,如深度(限价指令簿形状)是随时间变化而随机变化,投资者延迟交易还必须考虑流动性风险所带来的成本问题。比如,当流动性很高时,投资者应该加速交易,而当流动性较低时,投资者需要减少甚至等待交易。因

* 王丹、向修海,华中科技大学经济学院,邮政编码:430074,电子邮箱:wd5529@126.com。感谢匿名审稿人的宝贵意见,文责自负。

此，从随机流动性角度来看，最优执行策略的本质又是流动性风险控制问题，或者是随机流动性问题。

从以下的分析可以看出，最优执行问题建模的发展思路正好与对流动性设定的逐步完善相一致，从短暂效应到跨期效应，从确定性流动性到随机流动性。我们以这种思想将最优执行模型划分为三代：外生效应函数模型、内生反弹动态模型、随机流动性模型。接下来我们一一介绍其基本思想以及它们之间的区别和联系。

三、第一代模型：外生效应函数模型

这代模型首先由 Bertsimas & Lo (1998)提出，他们将大额投资者的执行成本定义为由一系列小额交易所带来的交易成本的期望值，然后他们使用动态规划求解得到相应的最优交易策略。这个思想后被 Almgren(2003)、Almgren & Chriss(1999，2001)扩展，在目标函数中同时考虑投资者的风险厌恶和交易成本的方差。但基本思想仍然是外生假设大额交易者的交易策略会对市场价格产生长期和短暂两种逆向效应。这也是这代模型的核心思想，外生的价格效应是它区别于以下两代模型的重要特征。国内对大额投资者交易策略的研究主要基于这种框架。为了便于理解这代模型的思想，下面的讨论将分别讨论这种框架的离散和连续时间版本。

鉴于直观理解，我们首先从离散时间形式出发，然后取交易间隔时间趋于 0 得到连续版本。假设市场存在风险资产（下面以股票为例）和无风险两种资产。大额投资者（以下统称“投资者”）需要在未来 $T(T<\infty)$ 时间内变现和买入数量为 X_0 的股票。下面仅考虑 $X_0>0$，即卖出所持有的股票，对于购买策略具有相似讨论。将时间区间 $[0,T]$ 划分成长度为 $\tau=T/N$ 的等距小区间，即在交易之前已经限制了交易的时间计划。相应地，投资者将股票分割成小额形式 $\{\zeta_i\}_{i=1}^{N}$ 分别在时刻 $t_i=i\tau(i=1\cdots,N)$ 出手，即 $X_0=\sum_{i=1}^{N}\zeta_i$。记在 t_i 时刻手头还剩余的股票数量为 $X_{t_i}=X_0-\sum_{j=1}^{i-1}\zeta_j$ 或者 $\zeta_i=X_{t_i}-X_{t_{i+1}}$，$X_{T+}=0$。为了引入长期和短期价格效应，我们还需引入第 i 个区间的平均交易变化速度 $V_i=\zeta_i/\tau$。在此假设下，可假设受到交易逆向冲击的价格过程[①]，即长期价格效应价格过程和总的股票价格过程（包含短期效应）分别为：

$$S_{i+1}=S_i+\sigma\tau^{1/2}u_{i+1}-\tau g(V_{i+1})$$

$$=S_0+\sigma\tau^{1/2}\sum_{j=1}^{i+1}u_j-\tau\sum_{j=1}^{i+1}g(V_j)$$

$$=S_i^{'}-\tau\sum_{j=1}^{i+1}g(V_j) \tag{1}$$

$$\tilde{S}_{i+1}=S_i-h(V_i)\qquad i=1\cdots,N \tag{2}$$

式(1)中的 $S_i^{'}$ 设定了长期均衡价格的动态过程，它是由无漂移项的算术布朗运动所决定。在实际中，它仅仅理论上存在，完全由市场基本面所决定。一般地我们能够观测到的是 $\tilde{S}$，它包含了大额交易的长期和短期价格效应。另外，其中 σ 是传统意义下价格的波动；$g(V_{i+1})$ 反映了大额交易对市场供给与需求的长期均衡影响，通过过程 S 不断累积在区间 $[0,T]$ 对价格都产生影响，因此它是长期效应，它是交易速度的函数；而 $h(V_{i+1})$ 仅仅只对区间 $[t_{i+1},t_{i+1}+1/N]$ 的交易价格产生影响，在短时间内会衰减消退，不存在持久性，它是短期效应；u 是相互独立的随机过程。因此，大额投资者的财富和执行成本分别为：

$$R^{'}(X)=X_0S_0-\tau\sum_{i=1}^{N}X_ig(V_i)-\tau\sum_{i=1}^{N}V_ih(V_i)$$
$$+\sigma\tau^{-1/2}\sum_{i=1}^{N}X_iu_i \tag{3}$$

$$I^{'}(X)=X_0S_0-\sum_{i=1}^{N}\zeta_i\tilde{S}_i^{'}=\tau\sum_{i=1}^{N}X_ig(V_i)$$
$$+\tau\sum_{i=1}^{N}V_ih(V_i)-\sigma\tau^{-1/2}\sum_{i=1}^{N}X_iu_i \tag{4}$$

式(4)的等式右边前两项分别表示长期效应和短期效应成本，而第三项为价格波动带来的成本。Almgren(2003)和 Schied & Schoneborn(2009)考虑了上述框架的连续版本，即令 $\tau\to\infty$，可得连续交易策略过程为：$X_t=X_0-\int_0^tV(s)ds$，这意味着 X 是绝对连续的过程，且 $V_t=\dot{X}_t$。Cetin 等(2010)将其扩展到无界变差情形的变现策略。同时考虑初始的无风险资产过程为 y，与(3)、(4)式相对应的连续形式财富过程和执行成本分别为：

$$R(X)=y+X_0S_0-\int_0^TX_tg(\dot{X}_t)dt$$
$$-\int_0^T\dot{X}_th(\dot{X}_t)dt+\sigma\int_0^TX_td\,W_t \tag{5}$$

$$I(X)=y+X_0S_0-R(X) \tag{6}$$

那么，其期望和方差分别为：

$$E[C(X)]=\int_0^TX_tg(\dot{X}_t)dt+\int_0^T\dot{X}_th(\dot{X}_t)dt \tag{7}$$

$$Var(C(x)) = \int_0^T \sigma^2 X_t^2 dt \tag{8}$$

Almgren & Chriss(2001)、Almgren(2003)、仲黎明等(2005)、胡小平等(2008)考虑了均值－方差最优化问题,即:

$\min_X \{E[C(X)] + \lambda Var(C(x))\}$,其中,$\lambda$ 测度了投资者的风险偏好程度。当 $\lambda > 0$、$\lambda = 0$、$\lambda < 0$ 分别对应风险厌恶、中性、偏好。显然投资者的最优交易策略是 λ 的函数,即 $X^*(\lambda)$,进而 λ 会影响到有效前沿 $(Var(C(X^*(\lambda))), E[(C(X^*(\lambda)))])$。由于在非线性效应情况下无法得到显性解,Almgren & Chriss(2001)考虑了如下线性情况:假设 $g(V_i) = \theta V_i$,$h(V_i) = \vartheta sign(V_i) + \eta V_i$。在风险中性情况下,如果忽略短暂效应,最优的交易策略为 $\zeta_i^* = X_0/(N+1)$,即将总的股票资产平均分割成等额头寸在每个时刻成交。它是一个常数速度的静态交易方案,与其它参数无关,即市场价格变化等信息并不会影响交易策略。从第二代模型可以看到,这并不是这个模型的缺陷,主要与线性反应函数有关。非线性效用函数的讨论见 Almgren(2012),最优交易策略将不再是这种平凡的交易,而是与效应函数参数有关。

国外对这代模型其它方面的研究并不是很多,反而国内学者对此做了大量扩展。最早的文献是刘海龙等(2005)和仲黎明等(2005)假设没有价格效应的资产价格服从连续的算术布朗运动,研究了机构投资者的变现策略。这对于变现时间很短是一个合理的假设。但对于变现时间较长时,几何布朗运动更加具有准确性,这个问题并不是没有意义,刘富兵、刘海龙(2007)和唐文芳、杨招军(2008)证明了二者的变现策略存在一定差别。国外文献都只讨论了确定性流动性和固定交易时间的变现策略,现实情况是大额交易产生的价格冲击造成的流动性成本也依赖于市场其他的参与者,例如噪音交易者的交易积极性,而交易对手信息是随机的,因此流动性具有随机性,即瞬间的冲击具有随机性。吕宏生等(2006)和储小俊、刘思峰(2007)最早在最优变现策略中考虑到随机的短期冲击效应,弥补了这种方法的缺陷。谢盐等(2006)假设了 L 形状的日内流动性,结果变现策略也随之变化。最后,因为上述所有文献都是基于单个资产的变现策略,对于一个机构投资者来说,一般会考虑组合最优变现策略,张丽芳、刘海龙(2009)最早将其推广到组合最优调整问题,发现股票之间的相关性提供了另外一条降低流动性成本的渠道。对于该代模型的其它主要缺陷,我们将在介绍第二代模型后再集中比较讨论,以凸显后者的优势。

四、第二代模型:内生反弹动态模型

第二代模型首先由 Obizhaeva & Wang(2011)提出,与第一代模型相比,除了考虑交易产生的长期和短期价格效应以外,它主要强调市场供给与需求的跨期动态特征。由于交易执行成本是一系列交易成本之和,跨期效应将使得最优交易策略完全不同于以前的形式,最重要的是它更加与实证结果相符合。与第三部分讨论一样,我们首先从离散时间形式出发,然后讨论连续形式。

(一)离散时间形式的最优化策略

在给出投资者的最优化问题之前,我们有必要首先讨论限价指令簿的动态特征,因为这是第二代模型的基础研究框架。为了便于描述,我们以电子化交易平台为例,这种交易系统是指令驱动型。投资者可以提交市价指令或限价指令。前者是一种在当前市场最优可成交价位上买或卖一定数量的风险资产的指令,它直接消耗掉市场上的限价指令或反方向的市价指令。后者是在预设的某个价位上买卖一定数量的风险资产指令,它并不会马上被成交掉,而是等待与之相匹配的反向市价指令。卖方限价指令形成限价指令簿(简记"LOB")的上半部分,而买方形成下半部分,它们反映了当前市场供给与需求状况。下面我们仅仅考虑 LOB 的上半部分,即大额投资者只提交买入的市价指令,对于卖出指令有类似结论。设最低卖方限价指令价为 A_0,最高买方限价指令价为 B_0,二者差额为买卖价差 $s = A_0 - B_0$。为了讨论 LOB 的动态,我们在此引入中间价格 $V_0 = (A_0 + B_0)/2$,则 $A_0 = V_0 + s/2$。对于 LOB 的形状,我们仅考虑块状结构,即在任意价位 P 上等待成交的限价指令有 q 单位,或者 $qI_{\{P \geq A_t\}}$,一般形状的讨论见 Alfonsi et al(2010)、Alfonsi & Acevedo(2012)。

下面我们讨论当大额投资者提交 x_t 单位的市价指令时,LOB 形状的动态变化。与第一代模型一样,将时间区间 $[0, T]$ 划分成长度为 $\tau = T/N$ 的等距小区间,但假设投资者需要购买 X_0 单位的股票。

设没有交易影响的价格过程为 F_t，它可以是算术布朗运动。记 A_t、V_t 表示时刻 t 的实际最优价格和中间价格，即考虑了 t 以前交易对价格的影响。定义 $D_t = A_t - V_t - s/2$ 为由过去交易所产生的额外价差。对于一般情况，在 $t = t_n = n\tau$ 时刻，投资者提交 x_t 单位市价指令，将消耗从 A_t 到 $A_{t+} = A_t + D_{t+} - D_t = A_t + s/2 + D_{t+}$ 的股票，这里 D_{t+} 由 $(D_{t+} - D_t) * q = x_t$ 决定。

与第一代模型类似，我们将第 t 期价格效应 $D_{t+} - D_t = x_t/q$ 分解成两部分，第一部分为长期效应 λx_t，第二部分为短期效应 κ，但是并不会立即瞬间至零，而是以速度 ρ 衰减，即 t 期价格效应 $\lambda x_0 + x_0 \kappa e^{-\rho t}$，其中 $q = 1/(\lambda + \kappa)$。长期效应会随着时间累计，这相当于对中间价格产生相应影响，即：

$$V_t = F_t + \lambda(X_0 - X_t) = F_t + \lambda \sum_{i=1}^{n-1} x_i$$

其中 X_t 表示在 t 时刻还需要购买的股票数量，而实际价格过程为：

$$A_t = V_t + s/2 + \sum_{i=1}^{n-1} \kappa x_i e^{-\rho\tau(n-i)}$$

其平均执行成本为：

$c_n = (A_{t_n}) x_{t_n} + \frac{D_{t_{n+}} - D_{t_n}}{2} x_{t_n} = (A_{t_n}) x_{t_n} + \frac{1}{2q} x_{t_n}^2$，总的执行成本为 $C = \sum_{n=1}^{N} c_n$。最优执行问题为求解下列约束问题

$$L_0 = \min_{\{x_1, \cdots, x_N\}} E_0 [\sum_{n=1} [A_{t_n} + x_n/2q] x_n] \tag{9}$$

$$\text{s. t.} \quad A_{t_n} = F_{t_n} + \lambda(X_0 - X_{t_n}) + s/2 + \sum_{i=1}^{n-1} x_i \kappa e^{-\rho\tau(n-i)} \tag{10}$$

Obizhaeva & Wang(2011)使用动态规划得到其最优策略为：

$$x_n^* = -\frac{1}{2} \delta_{n+1} [D_{t_n} (1 - \beta_{n+1} e^{-\rho\tau} + 2\kappa \gamma_{n+1} e^{-2\rho\tau}) - X_{t_n} (\lambda + 2\alpha_{n+1} - \beta_{n+1} \kappa e^{-\rho\tau})] \tag{11}$$

$$x_N = X_N$$

其中 δ_{n+1}、α_{n+1}、β_{n+1}、γ_{n+1} 由递归方程得到，具体见 Obizhaeva & Wang(2011)文章的命题 1。令 $N \to \infty$，得到极限情形：

$$x_1 = \frac{X_0}{\rho T + 2}, \lim_{N\to\infty} x_n/(T/N) = \dot{X}_t = \frac{\rho X_0}{\rho T + 2}, x_T = \frac{X_0}{\rho T + 2} \tag{12}$$

如果我们假设 $q = 1/(2\lambda)$，$\rho \to \infty$，那么第一代模型的对应形式是第二代模型的特殊情形，则最优执行策略为 $x_i^* = X_0/(N+1)$，$i = 1, \cdots, N$。显然与式(11)或(12)中的策略存在很大差别，式(12)中的策略是在开始和期末都有一个大的交易，中间的交易量较少。这是因为 LOB 的价格效应存在指数形式的反弹，反弹使得 LOB 动态存在跨期效应，投资者在中间时刻显然不能过多执行交易，但可以利用这个反弹获得一定的流动性以减少总的执行成本。而且式(12)显示在中间时刻是以固定速率的连续形式交易，因此第一代模型预先排除连续交易策略仅仅得到了次优的执行策略，同时说明交易的时间计划也很重要。下面将直接在连续时间框架下讨论离散和连续交易策略，以规避第一代模型的缺陷。

（二）连续时间形式的最优化策略

记连续交易策略为 μ_t，$t \in [0, T]$，μ_t 表示在时间区间 $[t, t+dt]$ 的交易数量。同时考虑离散形式的交易 $x_{\{t\in \hat{T}\}}$，其中 $\hat{T} = \{\tau_k : 0 \le \tau_k \le \tau_{k+1}, k \in N_+\}$ 这类似于脉冲响应控制问题。因此，总的交易策略空间为：$\Theta = \{\mu_{[0,T]}, x_{\{t\in\hat{T}\}} : \mu_t, x_t \ge 0, \int_0^T \mu_t dt + \sum_{t\in\hat{T}} x_t = X_0\}$，那么 X_t 的动态过程为：$X_t = X_0 - \int_0^t \mu_s ds - \sum_{s<t, s\in\hat{T}} x_S$ 参照离散形式，连续的实际价格过程为：$A_t = A_0 + \int_0^t (dV_S - \rho D_S ds - \kappa dX_S)$，其中 $A_0 = F_0 + s/2$，$V_t = F_t + \lambda(X_0 - X_t)$，$D_t = A_t - V_t - s/2$。

最后，相应的最优执行问题为：

$$L_t = \min_{\mu_{[t,T]}, x_{\{t\in\hat{T}\}}} E_t [\int_t^T A_S \mu_S ds + \sum_{t\le s\le T, s\in\hat{T}} [A_S + x_S/(2q)]] \tag{13}$$

其最优执行策略为(12)式，即由初始时刻和最终时刻离散交易外，其它中间时刻为连续交易策略。对上面基本模型的扩展有很多方面，主要针对 LOB 形状、反弹函数形式以及风险加权的均值方差问题等。首先，Alfonsi et al(2009) 扩展了 LOB 形状为非线性形式，通过分析发现最优化问题为一个确定性动态规划，通过解此规划得到了封闭型最优交易策略。从他们的分析可以看出，LOB 的形状直接决定了交易效应函数形式，块状的 LOB 形状对应于线性函数，而非线性 LOB 对应于非线性交易效应函

数。其次，Obizhaeva & Wang(2011)考虑了时变的指数反弹函数，结论仅仅是将 ρT 变成 $\int_0^T \rho_t dt$ 。Alfonsi et al(2009)考虑了一般反弹效应函数 $G(\cdot)$，如幂率反弹效应函数 $G(\cdot)=(1+\lambda t)^{-\gamma}$ 。但是在这方面的扩展并不是任意的，Gatheral(2010)证明不同价格效应函数与反弹效应函数的组合会导致价格操纵，或者在金融工程定价中称之为套利，这显然不是期望的结果。因为当存在套利时，投资者可以重复这个策略获得无穷大的期望收益。这既是不合法的，而且导致最优执行问题不存在解。Alfonsi et al(2010)证明非线性价格效应仅仅与指数反弹函数相容。Alfonsi et al(2010) 证明了当价格效应为线性形式时，反弹函数必须为严格凸函数，否则存在由交易触发的价格操纵。对于两个函数的一般形式组合结论仍然是悬而未解的问题。最后，对于考虑风险加权扩展问题，仅仅在 Obizhaeva & Wang(2011)原始论文中有所提及，其它文献都是在风险中性情况下考虑。主要原因是此扩展仅仅涉及到复杂的求解而并没有任何新的启发。但在下面的第三代模型中此扩展并不是没有意义的。

五、第三代模型：随机流动性模型

与前两代模型不同，随机流动性模型正处于模型构建初期，文献较少，据笔者所知，目前有吕宏生(2006)、储小俊和刘思峰(2007)、Almgren(2011)等文献，但这些文献都是在第一代模型基础上做出了扩展。在第二代模型基础上讨论随机流动性的文献仅仅有 Fruth et al(2011)，其在博士论文中对此做了全面的讨论。之所以将此扩展归为新一代模型，是因为流动性是最优执行问题的本质，并且实证上发现流动性是随机变化的，如见 Blais & Protter(2010)。当考虑随机流动性时，流动性风险就会影响最优执行策略。因此，最优执行策略的风险不仅涉及到价格风险，还与流动性风险有关。下面我们以 Fruth(2011)模型为例介绍这代模型。

从第二代模型可知，D_t 表示由过去交易所产生的额外价差。那么，假设在 t_n 时刻提交 x_n 市价指令，我们有 $D_t=\lambda\sum_{t_n<t}x_{t_n}+\sum_{t_n<t}k_{t_n}e^{-\int_{t_n}^{t}\rho_s ds}x_{t_n}$ ，即 D_t 以 ρ_t 速度均值复归，而且在每个交易时刻会有新的增加。与第二代模型的区别是我们将短期效应变成了时变形式，即 κ_t 。相应的连续版本为：$dD_t=-\rho_t(D_t-\lambda\Lambda_t)dt+(\lambda+\kappa_t)d\Lambda_t$ ，其中 Λ_t 为在 t 时刻投资者已经购买的股票数量。根据 Kyle(1985)对流动性的定义，ρ_t 、κ_t 分别表示流动性的反弹和深度两个要素。随机流动性是假设他们都是随机变化的，而 Fruth(2011)仅假设 κ_t 是随机的，譬如如下的扩散过程：$d\kappa_t=\alpha(t,\kappa_t)dt+\sigma(t,\kappa_t)dW_t$ 。在此记号下，二代模型中的式 $c_n=(A_{t_n})x_{t_n}+\frac{D_{t_{n+}}-D_{t_n}}{2}x_{t_n}=(A_{t_n})x_{t_n}+\frac{1}{2q}x_{t_n}^2$ 的连续形式为 $c_{dt}=(A_t+\frac{\Delta\Lambda_t}{2q})d\Lambda_t$ ；那么投资者的最优化问题为：$L_0=\min_{\Lambda_{T+}=X_0}E_0[\int_0^T(A_t+\frac{\Delta\Lambda_t}{2q})d\Lambda_t]$ ，其中 $A_t=A_t^u+D_t=F_t+s/2+D_t$ 。

不幸的是引入随机流动性后我们无法得到显性的最优交易策略。但能证明解的存在性和一些最优交易策略的性质，证明参见 Fruth(2011)的第 2.4 节。Fruth 证明最优值函数具有等待－购买结构(Wait region－ Buy region)，即在流动性高的时刻发出买方指令，而在流动性低时刻不进行任何交易。这就是引入随机流动性后导致结论与前面的完全不同。从上面的分析可以看到确定性情形仅仅是这代模型的特例，随机流动性的策略更加符合现实。尤其是在市场极端条件下，如金融危机时期，流动性变化显然不是确定的，极端变化的流动性显著地会影响到投资者的最优化策略。如果不考虑流动性的随机性，那么投资者会遭受巨大的执行损失。这代模型虽然具有一般性，但也有缺陷，一是前面提到的显性解不存在，二是实际应用比较困难。

六、未来发展评述与国内实践启发

前面我们归纳了最优执行成本理论的历史发展沿革，从中我们可以看到模型的设定不仅与市场的实际情况越来越相一致，而且模型逐渐完善，更加一般化。下面我们将评述最优执行成本理论的未来发展和对中国高频交易实践的启发。

(一)未来发展评述

通过比较三代最优执行成本理论模型，我们发现三代模型是相互包含的，并且第三代模型更一般、更与实际相符合。虽然第三代模型具有这些特点，但仍处于其发展初期，一些重要方面有待继续发展。笔者列举了如下几个重要问题，以促进理论发展和

实践应用：

1.价格效应函数的一般化。目前第三代模型还只是线性价格效应函数，从前面的分析可知这对应于 LOB 形状是块状的，即每个价位上等待买卖的股票数量是相等的。虽然第三代模型将其设定为时变的，不同时间发生变化，但在具体某个时刻，我们可以看到离最优价格越近，等待成交的限价指令数量越多。不仅这种实际情况需要考虑一般化，从理论方面来看，如果引入非线性价格效应函数，再加上随机流动性和一般反弹效应函数，那么最优化交易策略问题可能不存在，市场有可能存在价格操纵。所以，我们需要寻求条件使得模型的设定合理，并规避市场套利。

2.无统计或动态套利条件。在传统的定价文献中，只要存在鞅测度使得资产价格为鞅，那么市场是无套利的。其实这隐含了市场是无摩擦的，而我们的最优执行策略问题与之相反，考虑市场是有摩擦的，因此，需要发展原始条件。之所以这是最优执行策略问题的一个重要发展方向，是因为无套利定价涉及到或有权益的对冲问题，也就是一个特殊最优执行问题。这个问题也将成为未来发展的一个独立方向。研究的文献也逐渐增加，见 Jarrow(1992)、Elsinger & Summer(2001)、Alfonsi & Schied(2010)、Gatheral(2010)、Klock(2012)等。未来发展的基本思想是随机流动性过程以及随机反弹过程二者设定的一般化，当然要考虑到价格效应函数形式。

3.数值解技术。前面提及到第三代模型的缺陷是没有显性解，需要求助于数值解。如果数值求解 HJB 方程需要涉及到偏微分方程(PDE)，一般做法是有限元差分法。由解的结构可知，最优执行策略涉及到连续交易、离散交易以及无交易三种状态，显然这种方法存在较大误差，并不适合最优执行问题，因为最优执行一般是针对高量的交易，较小误差都会带来极大损失。因此，需要发展合理的数值技术，如马尔科夫值函数迭代方法等(Fruth，2011)。

4.交易的指令类型。前面三代模型都限定投资者只能提交市价指令。在流动性随机变化假定下，投资者显然可以利用限价指令获得更多机会。目前在这方面研究的缺乏主要原因是对机构投资者提交指令类型偏好的实证研究较少，没有典型特征可供建模分析。如果一旦成熟，这方面的研究将急需跟进。其发展方向是在前面的最优化问题中引入指令选择问题。

(二)国内实践启发

由于受到制度管制、交易速度和信息传递滞后，中国金融市场的高频交易并没有像西方发达市场盛行，在国外市场，据统计，高频交易占到了整个交易的 70%，而在国内仅仅在某些特殊市场上存在少数高频交易，如股指期货和商品期货市场。从前面对最优执行策略的研究来看，高频交易并不是一种非法交易，它的进入可以增加市场的活跃性，促进金融市场的发展和完善。因此，我们不应该抵制这种新型交易方式。并且只有为它提供发展空间才能使得国内的金融市场与国际接轨。下面我们将从最优执行这个角度讨论一些实践启发。

一是要加快政策法规改革。当前阻碍高频交易的政策法规过于严格，如高额的印花税和佣金、股市的“T+1”交易、卖空限制、换手率限制等。这些制度的初衷是稳定中国的金融市场，规避市场中的非法操作。从西方发达市场的发展过程来看，这些制度并不是监管者的最优选择，反而适当放开这些限制，增加市场的竞争程度，并通过提高市场的透明度，加强市场本身的监督才是发展出路。因此，从有利于整个金融市场发展的角度来看，这些人为的限制需要改革放宽。这样高频交易才能在中国市场实施，通过高频交易的正效应行为增加市场的竞争，从而反过来起到稳定市场的作用。

二是要推进 IT 产业的发展。首先是计算机的计算速度的提升，高频交易涉及到对未来价格、交易量、流动性等预测，然后最小化执行成本函数得到最优执行策略，这要求计算机在短时间内完成。因此，计算机的速度是最重要的。其次，计算机的稳定性也是很重要的，如果一旦程序错误将导致巨大损失。最后是通信技术的支撑，高频交易需要快速传递指令，并迅速返回成交信息(价格和交易量)，在发达市场上一般以微秒、纳秒为单位，而国内市场一般以秒为单位，如股票市场是三秒一笔明细。换句话说，高频交易需要 IT 产业的超前发展。

三是要重视相关专业人才的培养和建设。由于高频交易涉及到对市场变量的预测以及最优问题的求解，那么对高频交易的人才不仅需要经济学、金融学、统计学等发面的知识，还需要掌握计算机、数学等其它跨学科知识。因此，如果高频交易一旦实施，综合性人才的培养必须跟进。

综上分析，高频交易不仅有利于提高金融市场的流动性，同时对中国IT产业的发展、复合型人才的培养还具有带动作用。随着与国际市场的接轨，我们中国高频交易的蓬勃发展翘首以待。

注：

①例如，对于卖方交易策略，如果在当前市场价位上限价指令较少，那么大额交易者的卖方市价指令会使得最优价格下降。平均成交价并不是最优市价，而会低于该价。因此，交易前可观测的价格与交易后的实际价格有所区别。

参考文献：

Alfonsi, A. & A. Schied(2009), "Optimal execution and absence of price manipulations in limit order book models", Preprint.

Alfonsi, A. & J. Acevedo(2012), "Optimal execution and price manipulations in time－varying limit order books", Working Paper.

Alfonsi, A., A. Fruth & A. Schied(2010), "Optimal execution strategies in limit order books with general shape functions", Quantitative Finance 10(2):143－157.

Almgren, R. & N. Chriss(1999), "Value under liquidation", Risk 12:61－63.

Almgren, R. & N. Chriss(2001), "Optimal execution of portfolio transactions", Journal of Risk 3: 5－40.

Almgren, R.(2003), "Optimal execution with nonlinear impact functions and trading－enhanced risk", Applied Mathematical Finance 10(1):1－18.

Almgren, R.(2012), "Optimal trading with stochastic liquidity and volatility", SIAM J. Financial Math (1):163－181.

Bertsimas, D. & A. Lo(1998), "Optimal control of execution costs", Journal of Financial Markets 1(1):1－50.

Blais, M. & P. Protter(2010), "An analysis of the supply curve for liquidity risk through book data", International Journal of Theoretical and Applied Finance 13(6):821－838.

Cetin, U. R. Jarrow & P. Protter(2010), "Option hedging for small investors under liquidity costs", Finance and Stochastics 14(3):317－341.

Elsinger, H. & M. Summer(2001), "Arbitrage and optimal portfolio choice with financial constraints", Austrian Central Bank Working Paper No. 49.

Fruth, A. & T. Schoneborn(2011), "Optimal trade execution and price manipulation in order books with time varying liquidity", Quantitative Finance Working Paper.

Fruth, A.(2011), "Optimal order execution with stochastic liquidity", PhD Thesis, TU Berlin.

Gatheral, J.(2010), "No－dynamic－arbitrage and market impact", Quantitative Finance 10(7): 749－759.

Jarrow, R.(1992), "Market manipulation, bubbles, corners and short squeezes", The Journal of Financial and Quantitative Analysis 27(3):311－336.

Klock, F. & A. Schied(2012), "Price manipulation in a market impact model with dark pool", Working Paper, http://ssrn.com/abstract=1785409.

Kyle, A.(1986), "Continuous auctions and insider trading", Econometrica 53(6):1315－1335.

Obizhaeva, A. & J. Wang(2011), "Optimal trading strategy and supply/demand dynamics", Mathematics and Financial Economics 4:1－37.

Schied, A. & T. Schoneborn(2009), "Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets", Finance and Stochastics 13(2):181－204.

储小俊 刘思峰，2007：《在随机非线性价格冲击下的最优变现策略研究》，《中国管理科学》第5期。

胡小平 何建敏 吕宏生，2008：《非线性价格冲击函数下的最优变现策略》，《控制工程》第1期。

刘富兵 刘海龙，2007：《机构投资者变现成本的最优控制》，《上海交通大学学报》第12期。

刘海龙 仲黎明 吴冲锋，2005：《开放式基金流动性风险的最优控制》，《控制与决策》第2期。

吕宏生 胡小平 何建敏，2006：《随机冲击下的最优变现策略》，《系统工程》第12期。

谢盐 田澎 徐为山，2006：《基于日内流动性模式的最优变现策略》，《上海交通大学学报》第4期。

唐文芳 杨招军，2008：《机构投资者的最优变现策略》，《经济数学》第2期。

仲黎明 刘海龙 吴冲锋，2005：《机构投资者的最优变现策略》，《管理科学学报》第10期。

张丽芳 刘海龙，2009：《基于内生流动性风险的证券组合调整策略》，《管理工程学报》第3期。

（责任编辑：陈建青）