LLM 菜单定价

阳先毅

浙江大学计算机科学与技术学院

879946238@qq.com

2025 年 9 月 18 日

目录

相关参数刻画

任务：$i \in [0,1]$, 区间 $[0,1]$ 表示任务的索引。
模型投入部分：

- x_i: 任务 i 的输入 token。
- y_i: 任务 i 的输出 token。
- z: 用于微调 token，全局的。
- b: 不微调时模型的基础参数。

第 i 个任务的效用自然表示为：$v(x_i, y_i, z_i)$
自然的 assumption:

- $v_x(x,y,z) > 0, v_y(x,y,z) > 0, v_z(x,y,z) > 0$,concave。
- $v_{xy}(x_i,y_i,z) > 0, v_{yz}(x_i,y_i,z) > 0, v_{xz}(x_i,y_i,z) > 0$。
- $v(0,y,z) = 0, v(x,0,z) = 0$ 同时 $v_x(0,y,z) = v_y(x,0,z) = +\infty$
- $v_x(+\infty,y,z) = v_y(x,+\infty,z) = v_z(x,y,+\infty) = 0$

生产函数以及异质性 task 刻画

本文主要用科布道格拉斯函数作为生产函数：

$$v(x, y, z) = x^{\alpha} y^{\beta} (b + z)^{\gamma}$$

其中 $\alpha + \beta + \gamma < 1$, 这个时候我们就可以发现 b 的意义：即使模型不进行微调，只要有输入和输入也有一定的效用。

买方的任务可能是异质的。其边际价值由买方类型 $w = (w_i)_{i \in [0,1]}$ 表示，即对每个不同任务 i 的支付意愿：

$$w : [0, 1] \to \mathbb{R}_+$$

无限维太难了，有时做出简化（降为 2 维）：

$$w_i = \begin{cases} w, & \text{if } i \leq s, \\ 0, & \text{if } i > s. \end{cases}$$

$w \in [0, 1]$ 和 $s \in [0, 1]$,s 即 scale, 构成了阶梯函数。

目标函数（最有效率的 token 分配）

目标函数：

$$\max_{(x_i,y_i)_{i\in[0,1]},\, z\geq 0} \int_0^1 \left(w_i v(x_i, y_i, z) - c_x x_i - c_y y_i\right) di - c_z z$$

对于 x_i, y_i, 最优时必然满足：

$$w_i v'_x(x_i, y_i, z) = c_x$$
$$w_i v'_y(x_i, y_i, z) = c_y$$

对于 z, 有两种情况：

$$\int_0^1 w_i v'_z(x_i, y_i, z)\, di = c_z. \quad \text{选择微调}$$

$$\int_0^1 w_i v'_z(x_i^0, y_i^0, 0)\, di < c_z, \quad \text{不进行微调}$$

目标函数（柯布-道格拉斯）

在科布道格拉斯生产函数下有更清晰地刻画：

$$\max_{x_i,y_i,z\geq 0} \int_0^1 \left(w_i x_i^{\alpha} y_i^{\beta} (b+z)^{\gamma} - c_x x_i - c_y y_i \right) di - c_z z$$

我们首先固定 $z \geq 0$, 可以求 FOC 得：

$$x_i^*(z) = \left(\frac{\alpha w_i (b+z)^{\gamma}}{c_x} \right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\beta c_x}{\alpha c_y} \right)^{\frac{\beta}{1-\alpha-\beta}},$$

$$y_i^*(z) = \left(\frac{\beta w_i (b+z)^{\gamma}}{c_y} \right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\alpha c_y}{\beta c_x} \right)^{\frac{\alpha}{1-\alpha-\beta}},$$

$$U_i^*(z) = (1-\alpha-\beta)(b+z)^{\frac{\gamma}{1-\alpha-\beta}} \left(\frac{\alpha}{c_x} \right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y} \right)^{\frac{\beta}{1-\alpha-\beta}} w_i^{\frac{1}{1-\alpha-\beta}}.$$

目标函数（柯布-道格拉斯）续

总的：

$$X^*(z) = \left(\frac{\alpha(b+z)^\gamma}{c_x}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\beta c_x}{\alpha c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di,$$

$$Y^*(z) = \left(\frac{\beta(b+z)^\gamma}{c_y}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\alpha c_y}{\beta c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di,$$

$$U^*(z) = (1-\alpha-\beta)(b+z)^{\frac{\gamma}{1-\alpha-\beta}} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di.$$

我们发现固定 z 时，以上三式都有共同成分，记为：

$$\theta = \left(\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di\right)^{1-\alpha-\beta}.$$

我们可以算出 z 的角点解的临界条件，即对下式一阶条件恰在 z=0 时取到：

$$U^*(z) - c_z z$$

目标函数（柯布-道格拉斯）续

此时对应的 θ 恰为：

$$\hat{\theta} = b^{1-\alpha-\beta-\gamma} \left(\frac{c_x}{\alpha}\right)^{\alpha} \left(\frac{c_y}{\beta}\right)^{\beta} \left(\frac{c_z}{\gamma}\right)^{1-\alpha-\beta}.$$

If $\theta \leq \hat{\theta}$, then

$$x_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \frac{\alpha}{c_x} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}}, \tag{46}$$

$$y_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \frac{\beta}{c_y} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}}, \tag{47}$$

$$z^* = 0. \tag{48}$$

If $\theta > \hat{\theta}$, then

$$x_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \theta^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \frac{\alpha}{c_x} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}}, \tag{49}$$

$$y_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \theta^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \frac{\beta}{c_y} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}}, \tag{50}$$

$$z^* = \theta^{\frac{1}{1-\alpha-\beta-\gamma}} \frac{\gamma}{c_z} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}} - b. \tag{51}$$

代表型 θ

在柯布-道格拉斯生产函数和有效配置下，所有具有相同 θ 的类型消耗相同数量的微调 token、相同总量的输入与输出 token，并获得相同的总收益。任务 i 的最优输入与输出令牌数量与 $w_i^{\frac{1}{1-\alpha-\beta}}$ 成正比。类型进行微调当且仅当 $\theta > \hat{\theta}$。

因此，我们将 θ 称为代表性类型。在**价值-规模**设定下，每种类型 (w, s) 对应一个代表性类型：

$$\theta = \left(sw^{\frac{1}{1-\alpha-\beta}}\right)^{1-\alpha-\beta} = ws^{1-\alpha-\beta}.$$

菜单设计

我们假设买家的类型 w 满足分布 F（CDF），其中买家知道自己的类型，卖家只知道分布，我们希望设计一个直接机制，该机制指定了 token 在不同任务上的分配方案：

$$\big((x_i(w), y_i(w))_{i\in[0,1]}, z(w), T(w)\big)_w.$$

我们要注意到分配是可契约的（contractible）：卖家可以核实买家是否在相应任务分配了这么多代币。由于这是一个无限维筛选问题，难以求解，所以文章在以下两类（特殊情况）进行求解：

- 两类型情况。　　（次要）
- 价值-规模异质型。　　（主要）

two-types

假定只有两种类型 w_1 和 w_2，其先验概率分别为 f_1 和 f_2。此时，最优机制取决于：若卖方试图提取第一消费者剩余（即，提供一个菜单，其中包含每种类型的最优令牌数量，且价格等于其各自增加的价值），会有两种情况：

- 若此菜单是激励相容的，则它显然是最优的，我们将支付较高的类型标记为 H。
- 若此菜单不是激励相容的，则我们将激励约束被违反的类型标记为 H，另一类型标记为 L。

命题 (两类情形)

在最优菜单中，要么 *(i)* 卖方提取了全部剩余，要么 *(ii)* token 分配相对于虚拟类型 $\phi(w_H) = w_H$ 和 $\phi(w_L) = w_L - \frac{f_H}{f_L}(w_H - w_L)$ 是有效率的。

原文作者引用了Haghpanah and Siegel (2024)结论，一般筛选问题中两类买方紧约束结构的最新结果表明，在最优菜单中：

- 要么 (i) 卖方提取全部剩余；
- 要么 (ii) 类型 w_H 的激励约束和类型 w_L 的个人理性约束起作用。

无论哪种情况，类型 w_H 都会获得有效率的分配。
记 $q = (q_i)_{i=0}^1$ 为向每个任务提供的“质量”配置，其中 $q_i \triangleq v(x_i, y_i, z)$；并记 $C(q)$ 为生成给定配置 q 的最小总成本：

$$C(q) \triangleq \min_{x_i, y_i, z \geq 0} \int_0^1 (c_x x_i + c_y y_i + c_z z)\, di,$$

$$\text{s.t.} \quad v(x_i, y_i, z) = q_i, \quad \forall i \in [0, 1].$$

当不能提取全部剩余时，根据前人经验，有：

$$\max_{q_L, q_H, T_L, T_H} \quad f_L(T_L - C(q_L)) + f_H(T_H - C(q_H))$$

$$\text{s.t.} \quad \int_0^1 w_{Hi} q_{Hi} di - T_H = \int_0^1 w_{Hi} q_{Li} di - T_L, \qquad (\text{IC}_H)$$

$$\int_0^1 w_{Li} q_{Li} di - T_L = 0. \qquad (\text{IR}_L)$$

从两条约束把 T_H 和 T_L 换掉，有

$$\max_{q_L, q_H} \quad f_L \left(\int_0^1 \left(w_{Li} - \frac{f_H}{f_L}(w_{Hi} - w_{Li}) \right) q_{Li} di - C(q_L) \right) + f_H \left(\int_0^1 w_{Hi} q_{Hi} di - C(q_H) \right).$$

这就完成了证明。

two-types-科布道格拉斯

其实 w_H 和 w_L 是不容易判断的，因为他们是无限维的，但是在科布道格拉斯生产函数刻画下易于描述：

命题 (柯布-道格拉斯生产函数)

在柯布-道格拉斯生产函数的情形下，类型 w_H 是聚合指数 θ 更高的类型，即 $\theta_H \geq \theta_L$，或等价地：

$$\int_0^1 w_{i,H}^{\frac{1}{1-\alpha-\beta}} di \geq \int_0^1 w_{i,L}^{\frac{1}{1-\alpha-\beta}} di.$$

此外，最优菜单能够提取全部剩余当且仅当：

$$\int_0^1 (w_{i,H} - w_{i,L}) w_{i,L}^{\frac{\alpha+\beta}{1-\alpha-\beta}} di \leq 0.$$

two-types-科布道格拉斯

偏离时的效用：

if $\tilde{\theta} < \hat{\theta}$,

$$u(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \int_0^1 \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}} \; w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} di, \tag{52}$$

if $\tilde{\theta} \geq \hat{\theta}$,

$$u(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \int_0^1 \tilde{\theta}^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}} w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} di. \tag{53}$$

激励相容只可能在以下条件被打破：

$$\int_0^1 \left(w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} - \tilde{w}_i^{\frac{1}{1-\alpha-\beta}} \right) di > 0.$$

价值-规模异质性

$$w_i = \begin{cases} w, & \text{if } i \leq s, \\ 0, & \text{if } i > s, \end{cases}$$

w 和 s 独立，分别满足分布 F_w 和 F_s。买家的效用即：

$$w \int_0^s v(x_i, y_i, z)\, di - T.$$

我们记 $\int_0^s v(x_i, y_i, z)$，表示卖家应该向买家承诺的总质量水平，于是：

$$wq - T$$

卖家应该以最低的成本实现这个 q，即

$$C(q, s) = \min_{(x_i, y_i)_{i \in [0,s]}, z \geq 0} \int_0^s (c_x x_i + c_y y_i)\, di + c_z z \tag{1}$$

$$\text{s.t.} \quad \int_0^s v(x_i, y_i, z)\, di = q.$$

价值-规模异质性 (续)

因为 v 是严格凹的，所以最佳分配方法自然是将 token 平均分配到每个任务中：

$$C(q,s) = \min_{x,y,z\geq 0} \quad sc_x x + sc_y y + c_z z$$

$$\text{s.t.} \quad sv(x,y,z) = q.$$

有以下三个性质：

1. $C(q,s)$ 在 q 上是严格递增且严格凸的，且满足 $C_q(0,s) = 0$。
2. $C(q,s)$ 在 s 上是严格递减的。
3. $C(q,s)$ 是次模的，即对所有 q，其关于 q 的偏导数 $C_q(q,s)$ 随 s 增大而减小。

这 3 个性质都可以从 $v(x,y,z)$ 是严格凹的入手理解（证明略）。

续

激励相容要求：真实申报 w 比申报任意 $\tilde{w}$ 更优：

$$U(w) \geq w\, q(\tilde{w}) - T(\tilde{w}), \quad \forall \tilde{w}.$$

类似迈尔森引理，在临近 $\tilde{w}$ 处分析可得：

$$\frac{dU}{dw}(w) = q(w).$$

有

$$U(w) = \int_0^w q(k)\, dk.$$

对应的最优转移支付为：

$$T(w, s) = w\, q(w, s) - \int_0^w q(k, s)\, dk. \tag{2}$$

但是我们还没有确定 q。

续

卖方考虑最大化买家剩余：

$$\begin{aligned}\Pi &= \int \big(T(w) - C(q(w))\big) f(w)\, dw. \\ &= \int \left(wq(w) - \int_0^w q(k)\, dk - C(q(w)) \right) f(w)\, dw.\end{aligned}$$

经过变换积分次序等化简，有

$$\begin{aligned}\Pi &= \int \big(wq(w)f(w) - q(w)\big(1 - F(w)\big) - C(q(w))f(w)\big)\ dw. \\ &= \int f(w) \left[\left(w - \frac{1 - F(w)}{f(w)} \right) q(w) - C(q(w)) \right] dw.\end{aligned}$$

记 $\phi(w) = w - \frac{1-F(w)}{f(w)}$，我们知道一阶条件为

$$\phi(w) = C_q(q(w, s), s). \tag{3}$$

最优菜单的提出

假设：租金增长有限

对所有 w, s，由 (3) 式定义的函数 $q(w, s)$ 满足：

$$\int_0^w s\, q_s(k, s)\, dk \leq w\, q(w, s).$$

命题 (最优令牌分配菜单)

在假设下，一个最优菜单为：

$$\Big((x_i(w, s), y_i(w, s))_{i \in [0,s]},\ z(w, s),\ T(w, s) \Big)_{(w,s)},$$

其中，对每个 (w, s)：

- $(x_i(w, s), y_i(w, s), z(w, s))$ 是问题 *(1)*（即式 *(3)*）中定义的、以最小成本提供质量 $q(w, s)$ 的最优令牌配置；
- $T(w, s)$ 由式 *(2)* 定义。

最优菜单-proof

我们来证明买家没有偏离诚实报 (w,s) 的动机。首先说明买家不会报 $\tilde{s} < s$，当买家报 $(w, \tilde{s})$ 时，它的收益就是 $U(w, \tilde{s})$。由前面结论我们知道，$C_{qs} < 0$。在 (3) 中，我们对等式两边求导，有

$$\frac{\partial q}{\partial s} = -\frac{C_{qs}(q, s)}{C_{qq}(q, s)}.$$

C(q,s) 关于 q 是严格凸的，所以 $C_{qq} > 0$, 所以 $q(w, s)$ 关于 s 是严格单调递增的，因此

$$U(w, s) = \int_0^w q(k, s)\, dk.$$

随 s 而递增。因此，类型 (w, s) 不愿偏离至 $(w, \tilde{s})$，其中 $\tilde{s} \leq s$。另外，根据给定激励相容，$(w, \tilde{s})$ 也不愿偏离至 $(\tilde{w}, \tilde{s})$。因此，类型 (w, s) 不愿偏离至任何满足 $\tilde{s} \leq s$ 的 $(\tilde{w}, \tilde{s})$。

最优菜单-proof（续）

然后我们考虑买家报价 $\tilde{s} > s$ 的情况。若类型 (w, s) 偏离至 $(\tilde{w}, \tilde{s})$ 且 $\tilde{s} > s$，则其获得的总效用为 $w\, q(\tilde{w}, \tilde{s}) \cdot \frac{s}{\tilde{s}}$，并支付转移支付 $T(\tilde{w}, \tilde{s})$。因此，一个虚报类型的最优双重偏离策略由以下问题给出：

$$\max_{\tilde{w}} \left[w\, q(\tilde{w}, \tilde{s}) \cdot \frac{s}{\tilde{s}} - \tilde{w} q(\tilde{w}, \tilde{s}) + \int_0^{\tilde{w}} q(k, \tilde{s})\, dk \right].$$

我们发现上式实际上是类型为 $(w\frac{s}{\tilde{s}}, \tilde{s})$,$\tilde{s}$ 不偏报而 $w\frac{s}{\tilde{s}}$，但我们注意到当给定 $\tilde{s}$ 时，诚实占优，所以报的最佳的 w 就是：

$$\tilde{w}^* = w\frac{s}{\tilde{s}},$$

因此 (w,s) 偏报 $(\tilde{w}, \tilde{s})$ 获得效用为：

$$U(w; s, \tilde{s}) = \int_0^{w\frac{s}{\tilde{s}}} q(k, \tilde{s})\, dk.$$

最优菜单-proof（续）

我们对 $U(w; s, \tilde{s})$ 求关于 $\tilde{s}$ 的偏导，有：

$$\begin{aligned} U_{\tilde{s}}(w; s, \tilde{s}) &= \int_0^{w\frac{s}{\tilde{s}}} q_s(k, \tilde{s})\, dk - \frac{ws}{\tilde{s}^2} q\left(\frac{ws}{\tilde{s}}, \tilde{s}\right) \\ &= \frac{1}{\tilde{s}} \left[\tilde{s} \int_0^{w\frac{s}{\tilde{s}}} q_s(k, \tilde{s})\, dk - \frac{ws}{\tilde{s}} q\left(\frac{ws}{\tilde{s}}, \tilde{s}\right) \right] \leq 0, \end{aligned}$$

此处用到了前面的 assumption:

$$\int_0^w s\, q_s(k, s)\, dk \leq w\, q(w, s).$$

以上就说明了买家不会偏移到 $\tilde{s} > s$。

token 包套餐

在许多情况下，合同只能基于使用的 token 总数，而无法规定其在不同任务间的具体分配。为应对这一情形，作者允许卖方仅就输入、输出和微调 tojen 的总量进行签约，即出售“token 包套餐”。对应的菜单为：

$$(X(w),\, Y(w),\, Z(w),\, t(w))_w,$$

其中 $X(w)$、$Y(w)$ 和 $Z(w)$ 分别表示出售的输入、输出和微调令牌的总量。买方在购买后，可自由将输入和输出令牌分配至各项任务。同样，用科布道格拉斯生产函数表示如下：

$$\begin{aligned} U(X,\, Y,\, Z) = \max_{(x_i, y_i)_{i\in[0,1]}} & \int_0^1 w_i\, x_i^{\alpha}\, y_i^{\beta} (b + Z)^{\gamma}\, di, \\ \text{s.t.} \quad & \int_0^1 x_i di = X, \\ & \int_0^1 y_i di = Y. \end{aligned}$$

token 包套餐

对于这个凹的目标函数，我们用拉格朗日乘数法求解最值：

$$\mathcal{L} = \int_0^1 w_i \, x_i^{\alpha} \, y_i^{\beta} (b + Z)^{\gamma} \, di - \lambda \left(\int_0^1 x_i di - X \right) - \gamma \left(\int_0^1 y_i di - Y \right).$$

得到一阶条件：

$$w_i \alpha x_i^{\alpha - 1} y_i^{\beta} (b + Z)^{\gamma} = \lambda,$$
$$w_i \beta x_i^{\alpha} y_i^{\beta - 1} (b + Z)^{\gamma} = \gamma.$$

可以发现最值处 x_i, y_i 构成比值：

$$y_i = x_i \cdot \frac{Y}{X}.$$

带入到关于 x_i 的一阶条件当中：

$$w_i \alpha x_i^{\alpha + \beta - 1} \left(\frac{X}{Y} \right)^{\beta} (b + Z)^{\gamma} = \lambda.$$

再代到约束中表示出 x_i 为：$x_i = \dfrac{w_i^{\frac{1}{1-\alpha-\beta}}}{\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} \, di} X,$

接上一页

$$U(X,\,Y,\,Z)=\left(\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}}\,di\right)^{1-\alpha-\beta} X^{\alpha}\,Y^{\beta}(b+Z)^{\gamma}=\theta X^{\alpha}\,Y^{\beta}(b+Z)^{\gamma},$$

这里将 $\theta\in[0,1]$ 记为代表型，很合理，因为相同 θ 的买家会有相同的收益，因此我们又可以定义一个质量参数：

$$Q(X,\,Y,\,Z)=X^{\alpha}\,Y^{\beta}(b+Z)^{\gamma}$$

可以写出相应的优化问题：

$$\begin{aligned} C(Q)\triangleq \min_{X,Y,Z\geq 0}\quad & c_x X+c_y\,Y+c_z Z \\ \text{s.t.}\quad & X^{\alpha}\,Y^{\beta}(b+Z)^{\gamma}=Q. \end{aligned} \tag{4}$$

继续接上一页

买家最大化消费者剩余的过程和前面那个问题类似。成本函数 $C(Q)$ 是严格递增且严格凸的，且满足 $C'(0) = 0$。记 θ 的先验分布为 F_θ，定义虚拟类型：$\varphi(\theta) \triangleq \theta - \frac{1-F_\theta(\theta)}{f_\theta(\theta)}$，最终我们可以得到 Q 的决定式：

$$\varphi(\theta) = C'(Q(\theta)) \tag{5}$$

以及最优转移支付：

$$T(\theta) = \theta Q(\theta) - \int_0^\theta Q(k)\,dk. \tag{6}$$

命题 (最优令牌套餐菜单)

当只能对 *token* 进行签约，且虚拟类型 $\varphi(\theta)$ 递增时，最优菜单为：

$$(X(\theta),\, Y(\theta),\, Z(\theta),\, T(\theta))_\theta,$$

其中 $(X(\theta),\, Y(\theta),\, Z(\theta))$ 是成本最小化问题 *(4)* 中实现质量 $Q = Q(\theta)$ 的最优 *token* 组合，$Q(\theta)$ 由式 *(5)* 定义，而 $T(\theta)$ 由式 *(6)* 给出。所有对应于相同 θ 的原始类型 w 都会选择相同的菜单项。*(IC 已经保证不会偏离)*。

单维私有价值情况

假设买家具有单维类型 $\theta \in [0,1]$，虚拟类型 $\varphi(\theta) = \theta - \frac{1-F(\theta)}{f(\theta)}$ 递增，多维投入向量 $x \in \mathbb{R}_+^J$，线性投入成本函数 $\sum_{j=1}^J c_j x_j$，以及非线性估值函数：

$$\theta v(x_1, \ldots, x_J) - T,$$

其中 v 是严格凹且严格递增的。单位质量的最优成本函数：

$$C(q) \triangleq \min_{x_1,\ldots,x_J \geq 0} \sum_{j=1}^J c_j x_j,$$

$$\text{s.t.} \quad v(x_1, \ldots, x_J) = q.$$

生产函数严格凹，因此对应的成本函数 $C(q)$ 是严格凸的，消费者效用为：

$$\theta q - C(q).$$

再定义一个加价率为

$$m(\theta) \triangleq \frac{\theta}{\varphi(\theta)}$$

关键结论：两部定价可实现最优机制

引理

设 $(x_1(\theta), \ldots, x_J(\theta), T(\theta))_{\theta \in [0,1]}$ 为一个最优直接机制，对应的产出质量为：

$$q(\theta) = v(x_1(\theta), \ldots, x_J(\theta)).$$

则以下两部定价菜单（间接地）实现了该最优机制：

$$(p_1(\theta), \ldots, p_J(\theta), p_0(\theta))_{\theta \in [0,1]},$$

其中 $p_j(\theta)$ 是对投入 x_j 的线性价格，$p_0(\theta)$ 为一次性支付（固定费），其值为：

$$\begin{aligned} p_j(\theta) &= m(\theta)\, c_j \\ p_0(\theta) &= T(\theta) - m(\theta)\, C(q(\theta)). \end{aligned}$$

简要说明

对于给定的菜单项 $(p_1, \ldots, p_J, p_0)$，类型为 w（对应有效类型 θ）的买方问题可表述为“在支付最小化前提下的质量最大化”，即：

$$\max_{q \geq 0} \ \theta q - P(q),$$

其中

$$P(q) \triangleq \min_{x_1, \ldots, x_J \geq 0} \sum_{j=1}^{J} p_j x_j, \quad \text{s.t.} \quad v(x_1, \ldots, x_J) = q,$$

表示实现质量 q 所需的最小支付。
在所建议的定价方案下，有：

$$\sum_{j=1}^{J} p_j x_j = \sum_{j=1}^{J} m(\theta)\, c_j x_j = m(\theta) \sum_{j=1}^{J} c_j x_j.$$

因此，任何给定质量 q 的买方最优投入分配在成本意义上是有效的（即与成本最小化问题一致），且：

$$P(q) = m(\theta)\, C(q),$$

其中 $C(q) = \min_x \sum c_j x_j$ s.t. ,$v(x) = q$ 。

简要说明

买方最优的质量选择 $q(\theta)$ 由一阶条件决定：

$$\theta = P'(q) = m(\theta)C'(q) = \frac{\theta}{\varphi(\theta)}C'(q).$$

两边同时除以 $\theta > 0$，可得：

$$1 = \frac{1}{\varphi(\theta)}C'(q), \quad \Rightarrow \quad \varphi(\theta) = C'(q).$$

因此，买方选择的质量水平 $q(\theta)$ 恰好满足最优机制所要求的一阶条件。由于买家在给定 q 时优化投入与卖家优化成本是一致的，每个类型 θ 都会消费最优的投入组合 $x_1(\theta), \ldots, x_J(\theta)$。而 p_0 保障了在最优投入下买家刚好能获得信息租金 $\int_0^\theta q(k)\,dk$, 使得最后的收入能达到

$$T_\theta = \theta q^*(\theta) - \int_0^\theta q^*(k)\,dk$$

基本设定

我们现在将上述推理应用于价值-规模情形下完全可签约的 token 分配问题，并证明：可以通过一个带有任务规模 s 上限的两部定价菜单，实现可缔约任务基准中的最优机制。该菜单的形式为：

$$\left(p_x(w,s), p_y(w,s), p_z(w,s), p_0(w,s), s(w,s)\right)_{(w,s)},$$

其中，买方在选择某一菜单项后，需支付一笔预付费用 $p_0(w,s)$，并可以线性价格 $p_x(w,s)$、$p_y(w,s)$、$p_z(w,s)$ 自由购买输入、输出和微调代币，最多执行 $s(w,s)$ 项任务。
此时，相关的异质性为价值 w 的差异，相应的加价率为

$$m(w) \triangleq \frac{w}{\varphi(w)} = \frac{w}{w - \dfrac{1 - F_w(w)}{f_w(w)}}.$$

规模-价值双部收费菜单

租金增长有界

对于所有 w, s，由式 (3) 定义的分配 $q(w, s)$ 满足

$$\int_0^w q_s(k, s)\, dk \leq -\frac{w}{\varphi(w)} C'_s(q(w, s), s).$$

该假设可简单解释如下：对于任意类型 (w, s)，其边际租金的增加率不得超过该类型在生产其最优质量水平时的边际节省（或者说边际成本节省）。

命题

上述假设成立的条件下，以下形式为两部定价菜单：

$$\begin{aligned} p_j(w, s) &= m(w)\, c_j, \quad j = x, y, z \\ p_0(w, s) &= T(w, s) - m(w)\, C(q(w, s), s) \\ s(w, s) &= s, \end{aligned}$$

其中 $C(q(w, s), s)$ 由式 *(1)* 定义，$q(w, s)$ 和 $T(w, s)$ 分别由式 *(3)* 和式 *(2)* 定义，能够实现最优资源配置，并达到最优利润。

简要说明

对于两维类型的 $((w, s))$ 的买家，面对 token 价格 p_x, p_y, p_z，其最优代币分配问题为：

$$\max_{(x_i,y_i)_{i\in[0,s]},\, z\geq 0} \int_0^s \left(wv(x_i, y_i, z) - p_x x_i - p_y y_i\right) di - p_z z.$$

由于所有任务对称，买方会在每项任务中购买相同数量的输入和输出代币，因此该问题可重写为：

$$\max_{x,y,z\geq 0} \; s\left(wv(x, y, z) - p_x x - p_y y\right) - p_z z.$$

同样是在“支付最小的前提下实现质量最大化”

$$\max_{q\geq 0} \; wq - P(q, s),$$

其中

$$P(q, s) \triangleq \min_{x,y,z\geq 0} \; sp_x x + sp_y y + p_z z, \quad \text{s.t.} \quad v(x, y, z) = q/s.$$

简要说明-续

在命题 6 所给定价下，有 $P(q,s)=m(w)C(q,s)$，其中 $C(q,s)$ 是供给质量 q 的有效成本，。因此，若买方如实报告类型，则其消费的代币数量是有效率的，并支付最优的总转移金额。

$\rightarrow$ 对于偏离至 $(\tilde{w},s)$（即保持真实规模 s）的情况，类似前面的论证，此类偏离无利可图。

$\rightarrow$ 考虑偏离至 $(\tilde{w},\tilde{s})$ 且 $\tilde{s}<s$ 的情况。此时类型 (w,s) 获得与类型 $(w,\tilde{s})$ 相同的效用。同时，因为的信息租金 $U(w,s)=\int_0^w q(k,s)\,dk$ 关于 s 递增，该偏离即不会发生，而该条件成立。

$\rightarrow$ 考虑偏离至 $(\tilde{w},\tilde{s})$ 且 $\tilde{s}>s$ 的情况。由于最优加价率 $m(w)$ 不依赖于 s，必要且充分条件是预付费用 $p_0(w,s)$ 对所有 w 均关于 s 递增。由构造可知：

$$p_0(w,s)=wq(w,s)-\int_0^w q(k,s)\,dk-\frac{w}{\varphi(w)}C(q(w,s),s)$$

后一页我们就会发现 assumption 是怎么来的了。

续

$$\begin{aligned}\frac{\partial p_0(w,s)}{\partial s} &= w\frac{\partial q}{\partial s}(w,s) - \int_0^w \frac{\partial q}{\partial s}(k,s)\,dk - \frac{w}{\varphi(w)}\left(\frac{\partial C}{\partial q}(q(w,s),s)\frac{\partial q}{\partial s}(w,s) + \frac{\partial C}{\partial s}(q(w,s),s)\right)\\ &= -\int_0^w q_s(k,s)\,dk - m(w)\,C_s'(q(w,s),s)\end{aligned}$$

这不就正是我们前面的 assumption 吗？所以说确实有先射箭再画靶的味道了。

最优 token 包菜单

现在我们来制定 token 包套餐（即只约束求 X,Y,Z 的总量），可以设计一个**不设任务上限**的两部定价菜单，实现最优的套餐菜单。

$$(p_x(\theta), p_y(\theta), p_z(\theta), p_0(\theta))_\theta$$

其中，买方在选择某一菜单项后，需支付一笔预付费用 $p_0(\theta)$，并可按线性价格 $p_x(\theta)$、$p_y(\theta)$、$p_z(\theta)$ 自由购买输入、输出和微调代币，用于任意数量的任务。
θ 如第 7 页所定义。定义加价率：

$$m(\theta) \triangleq \frac{\theta}{\varphi(\theta)} = \frac{\theta}{\theta - \dfrac{1 - F_\theta(\theta)}{f_\theta(\theta)}},$$

最优 token 包菜单

命题

如果 $\varphi(\theta)$ 是递增的，则以下形式为两部定价菜单：

$$
\begin{aligned}
p_x(\theta) &= m(\theta)\, c_x, \\
p_y(\theta) &= m(\theta)\, c_y, \\
p_z(\theta) &= m(\theta)\, c_z, \\
p_0(\theta) &= T(\theta) - m(\theta)\, C(Q(\theta)),
\end{aligned}
$$

其中 $C(Q)$ 由式 *(4)* 定义，$Q(\theta)$ 由式 *(5)* 定义，$T(\theta)$ 由式 *(6)* 定义，能够实现代币套餐设定下的最优资源配置，并获得最优利润。所有对应于相同代表性类型 θ 的类型 (w, s) 都会选择相同的菜单项 *(*所以它才叫“代表型”*)*。

最优 token 包菜单

使用了“代表型”θ 之后，实际上 w 和 s 都被融入到了其中，即相同代表型的买家收益相同，表示如下：

$$\max_{\{x_i,y_i\}_{i\in[0,s]},\, z\geq 0} \left\{ \int_0^s \left[wv(x_i, y_i, z) - p_x x_i - p_y y_i \right] di - p_z z - p_0 \right\}$$

正好就是第 5 页的形式。（而我们已经证明代表型可以用到其中）
在所提出的定价方案下，有 $P(Q(\theta)) = m(\theta)C(Q(\theta))$，买方对代币的最优配置因此是有效率的，资源配置达到成本最小化。
根据 29 页的引理，该菜单的最优性直接成立。