Appendix

Why AMMs were created

AMMs were created to facilitate token trading on the novel Ethereum blockchain. Ethereum is an extension of the seminal Bitcoin blockchain in that it can execute simple programs and store relevant data and provide a store of wealth. Many contracts on Ethereum work together to create decentralized applications, aka 'dapps,' and these dapps are controlled by dapp equity token holders, who act like corporation shareholders. Blockchains have a native token used to pay its administrators (miners), and blockchains with smart contract ability can create an endless variety of tokens representing art (NFTs), dapp equity, and assets off the blockchain (e.g., wrapped bitcoin). Centralized exchanges like Coinbase and the Chicago Mercantile Exchange allow one to trade some tokens off the blockchain, like how one trades stocks on the NYSE, using a Central Limit Order Book (hereafter CLOB).

On a CLOB, market makers supply two-sided limit orders (i.e., bids and asks) and update their limit orders via cancelations and replacements thousands of times daily. These transactions are costless to send and can respond to market conditions within milliseconds via programs on servers co-located with the CLOB exchange database server. In contrast, blockchains have a thousand-fold greater latency and messaging costs of dollars as opposed to zero. Lastly, miners have control over the sequence of transactions within a block. With price-time priority generating huge rewards for being first, any profitable CLOB market maker would need an off-chain relationship with miners, creating myriad incentive problems and removing the transparency so valued by crypto users.

CLOBs are economically infeasible on a decentralized blockchain. The inherent latency and cost disadvantage of decentralized blockchains expose market makers to conspicuous adverse selection, creating large spreads that discourage liquidity traders.[1]

Basic AMM Mechanics

An AMM is an exchange enabled by blockchain technology where a smart contract calculates prices using an explicit formula referencing the contract's inventory of assets. The contract holds crypto assets (e.g., tokens, ether) and contains code defining how users can interact with it. Trades can happen as long as there are positive balances in the contract liquidity pool, without any interaction by others. This makes it like a vending machine on the blockchain, i.e., a smart contract. The pool token total and relative amounts and contract logic drive its behavior.

Liquidity Providers (LPs) deposit tokens into a trading pool in the contract, allowing traders to swap one token for another from the pool using a simple constant product formula. In the equation below, the product of the number of tokens A and B in the pool equals a constant k.[2]

$$A \cdot B = k \tag{0.1}$$

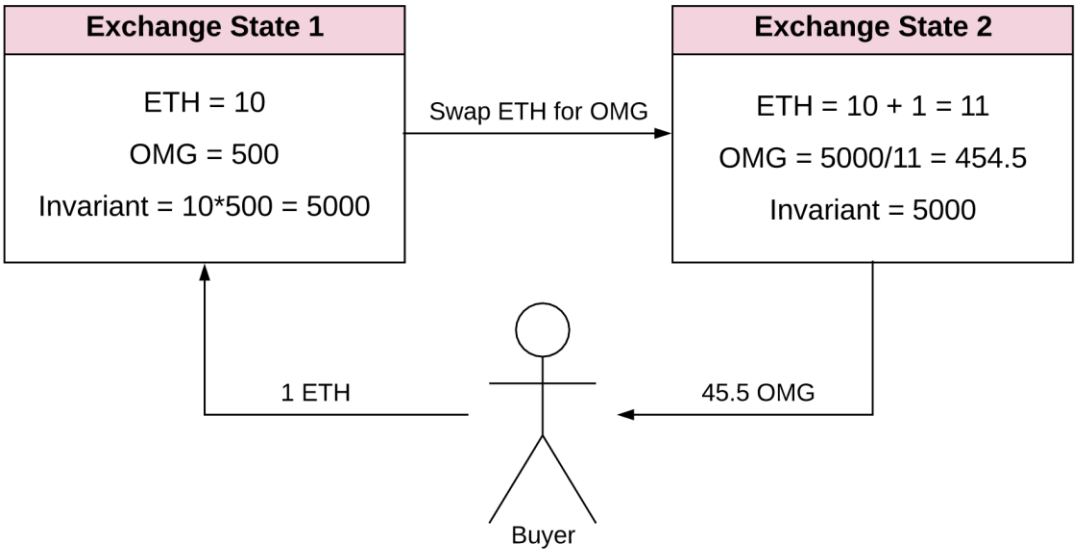
[1] See Appendix on Why latency kills CLOBs.

[2] There are also Constant Mean AMMs, which generalize to more than two assets, and generate the same problems addressed here as the basic two-token CPAMM, so that Constant Function AMM is a more general term.

n Figure 1, we see how a trader interacts with the popular Uniswap AMM, where k=5000.

Figure 1

ETH to OMG Exchange in Uniswap



In the following presentation, we will use the simple two-token constant product AMM because the results extend straightforwardly to multi-coin AMMs and refer to it simply as an AMM. In an AMM, the price of token A in terms of token B is determined by their relative quantities in the AMM's liquidity pool (hereafter, 'pool'). Without loss of generality, let us use ETH and a USD stablecoin token for tokens A and B, as it is more intuitive to think of an asset priced in terms of fiat currency, like a stock or commodity price. However, the mechanism can apply to pairs without a stablecoin, as in the Uniswap example above. The ratio of tokens in the pool determines the current price of one token in terms of the other:

$$price = \frac{USD^{pool}}{ETH^{pool}} \tag{0.2}$$

It is common to define the constant k as the square of the variable 'liquidity.' Using the square root of k, liquidity, as the metric of the LP's size generates a more intuitive metric of the size of the LP's position. For example, the pool's liquidity is the sum of the individual LP liquidity, which would not be the case for k.

$$k = liquidity^2 \tag{0.3}$$

Given this definition of pool liquidity and price, we can extend equation (0.1) to derive how transactions affect trades, prices, and pool token balances depending on what we consider the dependent variable. For example, given the AMM's liquidity and price, the pool amounts of USD and ETH.

$$\begin{aligned} USD^{pool} &= liquidity\sqrt{p} \\ ETH^{pool} &= \frac{liquidity}{\sqrt{p}} \end{aligned} \tag{0.4}$$

In equation (0.4) the priced asset pool quantity (here, ETH) is liquidity divided by the square root of price, and the unit of denomination in the pool (USD) is liquidity multiplied by the square root of price. Individual LP positions are captured by their liquidity, set at the time of deposit. The sum of individual LP pool balances and liquidity equals the total pool balances and liquidity.

Prices can be derived from relative token amounts in the contract and vice versa. Prices cannot change without a trade that changes a pool's relative quantities. LP actions—adding and removing liquidity—alter the depth of liquidity but not the AMM price or relative pool quantities. At the inception of a pool, an initial LP supplies tokens in a ratio consistent with the current price of the tokens. For example, an LP supplying 1600 USD tokens and 1 ETH token would imply the contract currently has an ETH price of $1600.

The contract's trade logic is based solely on enforcing that equation (0.1) holds after every contract interaction. These quantities refer to the pool, so positive amounts for ΔETH implies the trader is buying ETH, sending USD to the pool, and withdrawing ETH. For example, applying trades to equation (0.1) below, one can see how algebra can be applied to generate initial and ending price, or how one could calculate how much USD a trader will receive when he adds a given amount of ETH to the pool.

$$USD^{pool} \cdot ETH^{pool} = \left(USD^{pool} + \Delta USD^{pool}\right) \cdot \left(ETH^{pool} + \Delta ETH^{pool}\right) = liquidity^2 \tag{0.5}$$

Given the definitions for USD^{pool} and ETH^{pool}, and price, we can generate equations like the amount of ETH or USD sent from the trader to the pool as a function of liquidity, starting and ending price, p_0 and p_1, are as follows.

$$-\Delta ETH^{trader} = \Delta ETH^{pool} = liquidity \cdot \left(1/\sqrt{p_1} - 1/\sqrt{p_0} \right) \tag{0.6}$$

$$-\Delta USDC^{trader} = \Delta USDC^{pool} = liquidity \cdot \left(\sqrt{p_1} - \sqrt{p_0} \right) \tag{0.7}$$

The larger the trade size relative to the existing liquidity, the greater the price change.

$$\begin{aligned} p_0 &= \frac{USD^{pool}}{ETH^{pool}} \\ p_1 &= \frac{\Delta USD^{pool} + USD^{pool}}{\Delta ETH^{pool} + ETH^{pool}} \end{aligned} \tag{0.8}$$

Algebra gives various mathematically identical definitions of the post-trade price given the trade parameters and the initial price (e.g., as a function of the p_0, liquidity, and ΔUSD). The price at the transaction's start differs from the ending price since one token leaves and the other enters

the pool on any trade, changing the pool token ratio. The effective or fill price is between the two, specifically, the geometric mean of the start and end price.

The equations above allow us to estimate the critical cost borne by the LPs addressed in the LAMM.

AMMs are often referred to as part of 'defi,' which stands for 'decentralized finance,' in that it is controlled by a collective of people operating via pseudonymous accounts. These are 'trustless' programs in that no other agent has discretion over how the trade is processed, can change contract parameters, seize an account's tokens, or censor users from the contract. Smart contracts are administered by a collection of pseudonymous individuals on blockchains, which themselves are decentralized ledgers run over the internet worldwide. Ideally, AMMs are consistent with the crypto principles outlined in the Bitcoin White Paper (Nakamoto, 2008): immutability, censorship-proofness, permissionless, transparency, and anonymity. Many AMMs have piecemeal deviations from the ideal in various dimensions, such as sacrificing decentralization for speed. Nonetheless, AMM developers generally aspire to the Bitcoin principles.

Margin Trading Accounting Examples

Margined accounts imply that accounts can acquire negative net asset value. Any broker who gives customers margin needs a mechanism for liquidating accounts close to insolvency to protect themselves from losses that hurt their income and firm-wide insolvency.

The standard margin account gives customers leverage by allowing them to sell assets they do not have. Thus, with only ETH deposited, one can buy more ETH with USD one does not have and sell ETH with ETH one does not have. For example, the starting amounts on the left column represent two types of deposits, either in ETH or USD stablecoin. Both are worth $750. The ending accounts on the right represent two accounts with levered positions, both with a net asset value (hereafter, NAV) of $750, and the top is short ETH, the bottom long ETH.

ETH price $2000

USD deposit

item	quantity	$value
ETH	0.000	0
USD	750	750
ReqMargin		0
MktValue		750

Levered Short

item	quantity	$value
ETH	-1.000	-2,000
USD	2,750	2,750
ReqMargin		400
MktValue		750

ETH deposit

item	quantity	$value
ETH	0.375	750
USD	0	0
ReqMargin		150
MktValue		750

Levered Long

item	quantity	$value
ETH	1.000	2,000
USD	-1,250	-1,250
ReqMargin		400
MktValue		750

Either of the starting deposits on the left can generate long or short positions on the right using the following starting balances and trades:

- USD Deposit to Levered Short ETH
  - Sell 1 ETH, get 2000 USD
- USD Deposit to Levered Long ETH
  - Buy 1 ETH, pay 2000 USD
- ETH Deposit to Levered Short ETH
  - Sell 1.375 ETH, get 2750 USD
- ETH Deposit to Levered Long ETH
  - Buy 0.625 ETH, pay 1250 USD

Below are two accounts with a required margin ratio of 20%. The one on the left has no leverage, and the one on the right has leverage. The leveraged account is reflected in the USD negative balance, meaning the trader bought ETH with 500 USD he did not have. This −500 USD balance represents margin lending. Note the trader cannot withdraw all his ETH because that would violate his margin requirement. Thus, for the margined trader to withdraw his entire NAV, he must sell some of his ETH back to the pool.

Not Leveraged

ETH price:	2,000	
item	quantity	$value
ETH	1	2000
USD	2000	2000
ReqMargin		500
MktValue		4000

Leveraged

ETH price:	2,000	
item	quantity	$value
ETH	1	2000
USD	-500	-500
ReqMargin		500
MktValue		1500

For a short position, the contract provides margin lending allowing the trader to sell ETH he did not have. He must buy back his short ETH position to withdraw his entire balance.

Short Position

ETH price:	2,000	
item	quantity	$value
ETH	-1	-2000
USD	6000	6000
ReqMargin		500
MktValue		4000

The user who deposits ETH or deposits USD to buy ETH without leverage can withdraw his ETH position at any time and has no fear of liquidation if he keeps his ETH on the contract. He has a positive required margin, but as the value of his ETH goes to zero, his required margin goes to zero, so at any ETH price, his net asset value is always greater than his required margin.

He can withdraw the entire ETH position because the required margin constraint is evaluated using the portfolio balances after a withdrawal, which here would be zero. Posting USD, buying ETH, and later withdrawing USD would be like if he swapped USD for ETH without posting USD into the contract (i.e., just like a Uniswap swap of USD for ETH).

Non-Margined Long ETH or ETH Deposit

ETH price:	2,000	
item	quantity	$value
ETH	2	4000
USD	0	0
ReqMargin		1000
MktValue		4000

Trader Liquidation

In traditional markets, when the customer's net asset value is below their required margin ratio, they are usually notified, allowing them a day, or within the day, to liquidate their position at their discretion. If no action is taken, however, the broker will exit the position for the customer, often crudely generating additional losses via execution in a single large trade. The brokerage is more concerned about saving itself from the account's insolvency, so it does not mind giving away some money to market makers in a clumsy execution.

For AMMs, the liquidation process must be driven by an open process that incents outsiders to monitor accounts for margin violations and then liquidate them. There is no way for a smart contract to instigate a notification or liquidation automatically. By giving a profit when the trader breaches various criteria, that profit should predictably incent proper liquidations. The mechanism for liquidating a standard margin trading account could be as follows.

- Anyone with an active account can liquidate
- Liquidator's account acquires the defaulted trader's ETH position
- Liquidator's USD balance is debited from the ETH USD value
- Liquidator gets a fee

Liquidation would move the defaulter's ETH into the liquidator's account. The negative of the USD value of that ETH position is then credited to the liquidator, so the liquidator's immediate effect does not change his net asset value (NAV). A liquidator's fee would be in USD and calculated as a percent of notional that would be credited to the liquidator and debited from the defaulter's margin balance. The defaulter would pay a total fee split between the liquidator and the equity account.

Thus in the base case, we have the liquidator with ETH and USD balances denoted with the subscript q, a defaulter's balances with the subscript d, and an equity account with the subscript e.

	eth	usd
defaulter	ETH_d	USD_d
liquidator	ETH_q	USD_q
equity	ETH_e	USD_e

Suppose the liquidator takes out the defaulter when its margin ratio equals or exceeds the liquidation fee percentage. In that case, the defaulting account will have a positive NAV, and the reallocation of balances is as follows (here, p is the ETH price, and a fee is a number like 10%).

	eth	usd
defaulter	0	$USD_d-ETH_d*p-fee*abs(ETH_d*p)$
liquidator	ETH_q+ETH_d	$USD_q+ETH_d*p+fee/2*(ETH_d*p)$
equity	ETH_e	$USD_e+fee/2*abs(USD_d*p)$

The defaulter's account could be insolvent immediately after the liquidation, either because it had a negative NAV, or because their margin ratio was positive, but less than the total liquidation fee (e.g., the fee was 10% and their margin ratio was 8%). In this case, the defaulter's account is zeroed out (deleted), and the Equity balance's USD would be credited an amount so that all active accounts on the AMM have positive value.

If {NAV(defaulted Account) < 0}

Then {Equity USD Balance Credit: $USD_d – ETH_d*p – fee/2*abs(ETH_d*p)$}

This captures the case where the defaulter's NAV is negative after liquidation. The liquidators will be incentivized to liquidate insolvent accounts, as the equity account absorbs the losses from the defaulter's insolvency, including the fee required for the liquidator.

Potential equity account losses imply a positive equity account balance is needed to absorb potential losses. That is, the equity account balances act as an insurance fund. With an insurance fund all on contract, the LAMM can remain completely decentralized because no off-contract promise is involved. More importantly, it gives the equity owners of the LAMM an economic purpose for their equity tokens, just like the owner's equity on a corporate balance sheet: it is a cushion for losses.

The liquidator assumes the new position, which implies a liquidator needs sufficient NAV so that he is not instantly liquidated once he acquires the position. While that is a cost for the liquidator, it has the beneficial property of preventing cascading liquidations. The liquidator is incentivized to exit his new position efficiently instead of instantly to avoid losses generated by price impact. The liquidity fee paid to the liquidator should suffice for the cost of this capital requirement and provide the liquidator with sufficient profit to exit his acquired position without losing money on the liquidation.

Forcing the liquidators to acquire the defaulter's position imposes a risk on them, as in the case where prices are moving quickly. The liquidation fee should be sufficient to compensate for such

adverse movements, but not in all cases. A 5% fee would generate a loss to a liquidator who did not sell immediately a loss at most once a year. While it would be nice if liquidators had zero risk, there are costs to this—cascading liquidations—and as with everything, one must balance trade-offs. A liquidator can still expect to make money on average with high probability, and so a loss once every year or two due to market jumps should not dissuade liquidators.

Liquidators can acquire partial amounts of a defaulting account. This is useful because a significant position can imply a required margin the liquidator does not have. Several liquidators can acquire parts of the defaulter's position, or a single liquidator may complete the liquidation piecemeal.

To avoid manipulation, one can use a moving average instead of a current price. The trade-off here is that the longer the time dimension of the EMA, the slower the liquidation will be, subjecting the AMM to more insolvency risk. A balance must be made that protects accounts from manipulative liquidations vs. protecting the AMM from insolvency.

LP Capital Efficiency and Liquidation

The liquidation of an LP is different because, given leverage, their greater risk to the AMM is an absence of a token instead of insolvency. It is helpful to see why a leveraged AMM economizes on capital compared to a v3 LP position. First, consider the leverage LP account's balance sheet.

LP account at Initial Deposit

	LP Account	
	token	quantity
LP pool	ETHp	liq/sqrt(p)
	USDp	liq*sqrt(p)
Margin	ETHm	$-ETH_p*21/22$
	USDm	$-USD_p*18/19$
Net	ETH	$ETH_p/22$
	USD	$USD_p/19$

The figure above shows that the pool account balances are calculated in the standard way for an unrestricted range. The LP is given 22x leverage on his ETH so that 1/22 of his pool balance is his net ETH position, and 21/22 of his pool ETH position is reflected by a debt of 21/22 in his margin ETH balance. For the LP's USD, his pool position is leveraged 19 times. The asymmetry is applied because the LP loses USD quicker when prices rise than it loses ETH on price declines, and these numbers generate an approximate symmetry where the LP runs out of net USD or ETH on up and down 10% price movements.

To see how this affects the capital relative to v2 and v3 AMMs, we shall compare it to the v3 AMM, where the top and bottom range prices are 10% above and below the current price, as with a ±10% v3 range.

V2 vs. LAMM LP

StartPrice	1,500	
liquidity	1,000	

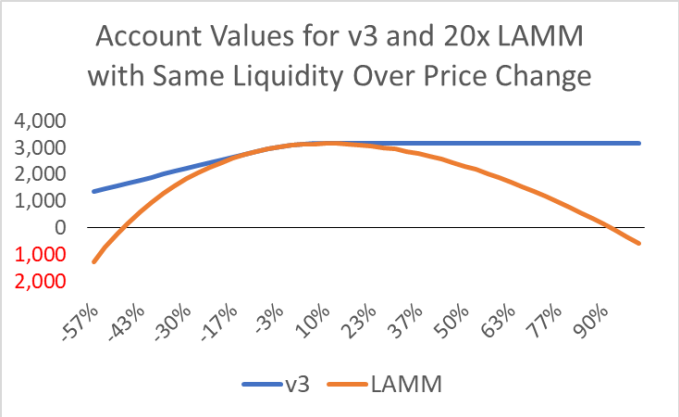
	v2	v3 +/- 10%	Levered v2
ETH	25.82	1.20	1.17
USDC	38,730	1,987	2,038
MktVal$	77,460	3,790	3,799
cap/v2Capital		4.9%	4.9%
v2Capital/cap		20.44	20.39

The leverage parameters are chosen above to generate virtually identical capital efficiency as the +/- 10% v3 range.

Thus, not only does a leveraged LP approach generate comparable capital efficiency to a v3 AMM, but it also facilitates the following

- A mechanism to remove IL
- Reduced CPU on contract
  - No looping through adjacent ranges with their different liquidity on single trades
- Reduced memory
  - No need to store tick data relevant to apportioning fees to LPs, which is about half of the state variable writes in v3
- Simplified LP aggregation making third-party vaults more viable

While the v3 LP can never default, the levered AMM LP can. However, even at 20x leverage, the price would need to move either down 50% or up 88% for this to happen (assuming the LP does not trade).



A trader with 20x leverage becomes insolvent after only a 5% price move. As the daily average ETH price volatility is 4%, this would generate a reckless probability of insolvency for LPs if they had the same risk for the same amount of leverage. However, the LP provides both ETH and USD in equal USD amounts at inception, unlike a trader whose leveraged or short positions are financed solely with USD. For example, a trader can invest 20 USD in financing a long position

of 100 USD worth of ETH. For an LP, he has 95 USD worth of ETH as a debt, reflected in a margin balance of -95, behind his 100 USD worth of ETH in the pool, reflected as a pool balance of +100.

If pool positions did not change with the price, the LP would have zero risk in his leveraged position, as both token margin debts are overcollateralized in the pool. This is another way to see that the IL is real. However, the IL is subtle, which is why a 20x LP leverage generates so much more risk than trader leverage, where 20x leverage is not subtle.

Given the AMM LP generates a negative token balance in one of the tokens more quickly than he becomes insolvent, and maintaining a positive token balance for each token in the pool is essential for a functioning AMM, LP liquidation focuses on this case as opposed to insolvency. The LP's primary job is to supply the contracts with sufficient USD and ETH. While an insolvency condition applied to the LP is still functional, this addresses unlikely edge cases.

On Uniswap v3, the restricted ranges create the possibility of running out of ETH or USD as the price moves. The profit incentive created by such a scenario lures in LPs, so that does not happen. For the levered AMM, there are two incentives. First, if the LP is hedging their position to remove their IL, they will not run out of either token, as they will be offsetting their pool token declines with net margin balance increases. This should be sufficient for most LPs.

For negligent or irrational LPs, liquidation is needed to prevent LPs from generating a prominent negative net token position that could leave the contract with insufficient tokens to allow swapping one token for another. Thus we need not have the liquidator acquire the LP's positions; transfer them from his pool position to his margin balance. The liquidator can apply the trader liquidation mechanism if the former LP violates his margin requirement as a trader after liquidation. Indeed, on the blockchain, this could happen within a single block.

For an LP liquidation, then, the rule is as follows. Given the LP has too much or too little of one of the tokens, he is then subject to LP liquidation. The liquidator then transfers the LP's pool balances to his margin balances, sets the LP's liquidity to zero, and takes a fee.

For example, consider the case where an LP starts with the standard leverage described above. He has negative margin balances reflecting how many tokens he borrowed to put into the pool.

LP on instantiation, ETH Price 1500

Liquidity = 852

Assets	quantity	$ value	Net	quantity	$ value
Pool	852		ETH	1.00	1,500
ETH	22.00	33,000	USD	1,737	1,737
USD	33,000	33,000			
Margin					
ETH	21.00	31,500			
USD	31,263	31,263			
Total Assets		3,237	Net Asset Value		3,237

Now assume the price of ETH fell, causing the LP's net ETH position to go below zero, which is the liquidation trigger.

LP after Price Change from 1500 to 1350

No trades in this account

Liquidity = 852

Assets	quantity	$ value	Net	quantity	$ value
Pool	852		ETH	2.23	3,001
ETH	23.23	31,246	USD	17	17
USD	31,246	31,246			
Margin					
ETH	21.00	28,245			
USD	31,263	31,263			
Total Assets		2,985	Net Asset Value		2,985

After liquidation, the LP's account would become a standard non-LP account and look like the account below [I am excluding the fee here, but it is a trivial debit from the ex-LP to the liquidator].

Ex-LP Account after Liquidation

Liquidity = 852, i.e., not an LP

Assets	quantity	$ value	Net	quantity	$ value
Pool	0		ETH	2.23	3,001
ETH	0	0	USD	17	17
USD	0	0			
Margin					
ETH	2.23	3,001			
USD	17	17			
Total Assets		2,985	Net Asset Value		2,985

LP Vaults

Given that the LPs in a leveraged contract only differ in their liquidity, providing a vault aggregating LP positions is much simpler for someone with a comparative advantage. With v3, the different ranges implied by LPs depositing at different times would generate complex optimal tactics. With the leveraged v2, the LPs have the same objective regardless of their initial price deposit. The vault manager would always be to arbitrage the AMM price with a centralized CLOB price based on their current and initial ETH net position. Facilitating such a service is valuable because while it is straightforward to automate an arbitrage algorithm, there are high fixed costs for those unfamiliar with programming. Many would be interested in providing liquidity but find creating a blockchain auto-bot beyond their capabilities. It is the sort of service suitable to a division of labor and subject to economies of scale.

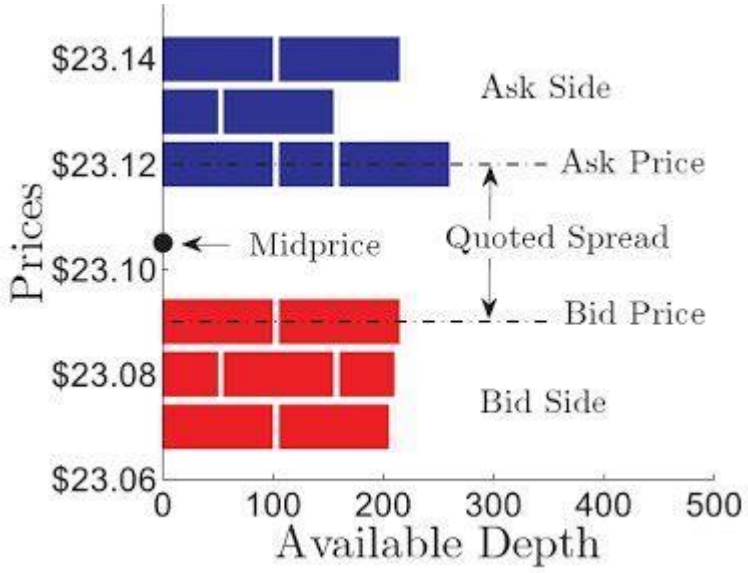
Economies of scale might sound like dangerous centralization, but as the contract takes in ETH and USD, outsiders can still arbitrage the AMM outside the fee (e.g., 0.3%) if it posts ridiculous prices. The vault manager still has the correct profit-maximizing incentive vis-à-vis the contract. Further, his customers should see centralization as an existential risk, as with mining farms on a blockchain. While a single agent may have the correct incentives, including creating an infrastructure diversified across servers worldwide, outsiders can never be sure there are no attack surfaces within his operation.

Prudent LPs should be wary of letting too few LP aggregator vaults administer the AMM to mitigate censorship or key-man risk. The key to incentivizing market competition is not the number of insiders but free entry by outsiders.[3]

Why Latency Kills CLOBs

Today's most efficient and liquid exchanges are central limit order books (CLOBs). It is a computerized system that aggregates and matches buy and sell orders for a particular financial asset, such as a stock, currency pair, or commodity, in real time. In a CLOB, market participants can place their orders to buy or sell a specific asset, and these orders are then listed in the book according to their price and time priority. The system automatically matches buyers and sellers at the best available price, filling limit orders that arrive first (i.e., price-time priority). Orders outside the current market price sit on the book as resting limit orders, available for traders until they are canceled.

A Limit Order Book



While these represent the gold standard, they require a degree of low latency (i.e., speedy) that is only attainable because they are centralized (ergo, CLOB and not LOB). This allows market makers to physically place servers running their market-making algorithms next to the exchange

[3] See the Theory of Contestible Markets

servers, allowing communications within milliseconds. Any decentralized trading platform cannot directly compete with these platforms as a price discovery mechanism.

To understand why a CLOB cannot work on a blockchain, it is helpful to understand the economics that drives its equilibrium. A CLOB has three types of traders: uninformed, informed, and market makers. The informed invest in data, models, and hardware to identify temporary mispricings and use their comparative advantage to snipe stale limit orders. The uninformed are ignorant for rational and irrational reasons: they may want to buy a car (liquidity traders) or are delusional and trading on irrelevant information (unintentional noise traders). As informed and uninformed traders do not show up simultaneously, market makers arise to provide liquidity continually by posting resting limit orders to buy and sell.

In equilibrium, all groups generate benefits equal to their costs. The informed trader's costs—investments in hardware and statistical algorithms—are balanced by revenue from adversely selecting stale market maker limit orders. Uninformed traders pay the market maker by crossing the spread, benefiting from convenient, quick trading. Market makers post knowing they will trade with both types of traders, setting limit orders such that the revenue from uninformed balances that they lose to the informed.

There are many scenarios where low latency is costly, but one applied to a market maker providing resting limit orders should suffice. A market maker places a two-sided order to buy or sell 100 shares of XYZ stock trading at a bid-ask price of $20.17-$20.18, its current bid price, $20.17. Assume the stock will move up or down $0.05 before you can cancel that order. If you get filled, it will only be because it is now trading at $20.12-$20.13, meaning you bought at $20.17 and can sell now at $20.12; you lost $0.05; if the price went up $0.05, you would probably not be filled on your $20.17 order to buy. This is called 'adverse selection,' where conditional upon getting filled, you paid too much or sold too low. It generates a loss profile for market makers like for those selling straddles or a liquidity provider's impermanent loss.

The effect of higher latency on a central limit order book is a classic example of Akerloff's 'Market for Lemons' (Quarterly Journal of Economics, 1970).[4] In that paper, he analyzes markets where parties with asymmetric information separate, so the only viable transactions are those with a negative value, and the market collapses (i.e., no trades).

The lemon's problem applied to limit order books is the following. High latency leads to market makers suffering higher adverse selection as it amplifies the relative speed advantage of informed traders, causing market makers to increase their spreads. Higher spreads discourage uninformed traders. With fewer uninformed traders, the market maker widens his bid-ask spread further to protect against adverse selection by the informed traders by making more profit per uninformed trade. This discourages more uninformed traders, creating a positive feedback loop until none are left.

[4] In this application, the market maker trading with informed traders generates a loss, like a lemon car, hoping to offset this with his trades with 'peach' cars (gains). In 70's slang, a 'lemon' is a bad type, as opposed to a good type, e.g., a 'peach.'

Another way to think about the necessity of liquidity traders focuses on the zero-sum nature of trading without them. With no liquidity traders, the remaining participants are then playing the unattractive game of trying to outsmart and out-speed others who have made the same commitment. The Milgrom-Stokey 'no-trade theorem' (Journal of Economic Theory, 1982) states that if all the traders in the market are rational, all the prices are rational. Thus anyone who makes an offer must have valuable and accurate private information, or else they would not be making the offer. Similarly, Grossman and Stiglitz's 'Impossibility of Informationally Efficient Markets' (American Economic Review, 1982) shows how without liquidity traders, no one has an incentive to put information into markets because the other rational traders infer what he knows via his market demand. There are no trades because every order is presumed to be informed and thus unprofitable for the other side.

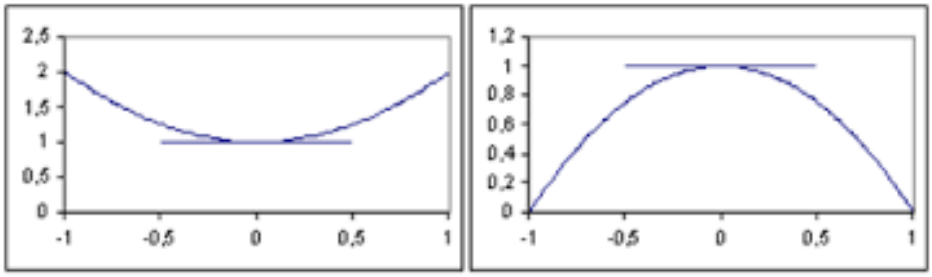
A deficiency of liquidity traders creates a positive feedback loop that causes markets to ultimately unravel. There will not be enough liquidity traders to support an active set of market makers, who need the uninformed retail flow to offset their losses to the informed traders. The high spreads and meager volume on decentralized limit order book exchanges are consistent with this result (e.g., Augur).

One can imagine the layer 2 blockchains will eventually become fast and secure, preventing this problem. However, even in this case, miners or validators can sequence transactions with some discretion. It takes 60 milliseconds for light to travel from Tokyo to San Francisco, creating a large lower bound to this discretionary time window. Successful market makers on modern CLOBs have reaction speeds of 5 milliseconds, implying the feasibility of front-running such a system with impunity (see Aquilina. Budish, and O'Neill, 2020).

A CLOB has price-time priority, so it fills limit orders first by price, and within a given price using first-in-first-out logic. Even without a minimum tick size, the sequencers could front-run limit orders by posting orders conditional upon the price in the new orders. There is no way for the layer 2 validators to agree on the time sequence of transactions if it is configured to prevent censorship, which would require a globally distributed set of validators. Given the disproportionate advantage of being first on limit order books, the unavoidable sequencing discretion makes transparent competition impossible, enabling and encouraging corruption.

Low-latency chains like Solana, meanwhile, are centralized, which invariably leads to corruption via Acton's Law. This centralization is not obvious, as many have more validators than Bitcoin or Ethereum (e.g., EOS has 21), but this Nakamoto coefficient is meaningless because the validators on low latency blockchains have to work together, and they are invariably controlled by a central agent. When a blockchain representative proclaims a bald-faced lie about a foundational crypto principle, its developers fall down the slippery slope, leading to more lying and, ultimately, a cesspool of deception. In markets dominated by unaccountable insiders, we should expect every sort of malicious trading tactic (e.g., FTX pumped its Serum token via its low-latency Project Serum exchange on Solana). This leaves blockchain CLOBs only for tokens with no alternatives, like in markets for NFTs and shitcoins.

Convexity Cost Inevitability



Convex function at a static point: local min

Concave function at a static point: local max

Convex payouts have a positive second derivative, while concave payouts have a negative second derivative. Concave payouts are often called negative convexity for this reason. Options have positive convexity for those who own them (aka, buy them, are long), so those who sell (aka 'short') options thus have negative convexity.

Jensen's inequality states that convex functions have 'time value' in that their expected value is worth more than their current 'intrinsic' value.

$$E\left[f\left(x\right)\right] > f\left(E\left[x\right]\right)$$

To see the inherent cost of negative convexity, consider a derivative valued at the square of the underlying price, p^2. This derivative would be an inefficient mechanism for generating convexity, which is why active options markets use strike prices instead of squared prices, but it makes for a simple example.

The delta of this security is its derivative, 2p, so a seller of this option would hedge it by going short 2p units of the underlying. If p=10, then the seller would hedge by shorting 20 units. The net payoff space would then be as follows:

				hedge	
	p	**p^2**	**p^2 pnl**	**pnl**	**Net**
initial	10	100	0	0	0
up	11	121	-21	20	-1
down	9	81	19	20	-1

The naked option seller is exposed to payouts of -21 and +19, while the hedged seller's payout volatility in the far right column is zero. It does not affect the expected value of the seller's payout.

In the Black-Scholes equation below, V is the option value, S is the stock price, r is the risk-free interest rate, and σ the volatility.

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0$$

The four terms in this equation can be interpreted as follows.

$$\frac{\partial V}{\partial t}=theta\ aka\ time\ decay$$
$$\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}=\frac{1}{2}\cdot variance\cdot gamma=convexity\ cost$$
$$r\left(S\frac{\partial V}{\partial S}-V\right)=\ financing\ cost$$

The equation shows the option's time decay (theta) plus its convexity return (gamma) equals the riskless return from a long position in the derivative and a short position and a hedged amount (*dV/dS* shares) of the underlying. Note that $\left(S\frac{\partial V}{\partial S}-V\right)$ is the cost of the initial position: the hedge value is the stock price times the option delta, $S\frac{\partial V}{\partial S}$, minus the option price, V.

We can see the intrinsic connection between theta and convexity. If we ignore the financing costs by assuming they are zero, we see that theta equals the negative value of the convexity, consistent with Jensen's inequality above. Theta, time decay, is required for an options market equilibrium because the convexity costs afflicting option sellers must be offset so that their net profit is zero, not negative.

$$\frac{\partial V}{\partial t}=-\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}$$

This derivation assumes transaction costs are zero, so it is fundamental. If hedging could remove hedging costs, options would have no time premium, but this never happens, highlighting the ineradicable nature of convexity costs. Convexity shifts the payoff space around; it doesn't reduce the expected cost of negative convexity.

Hedging is still recommended for option sellers because it decreases indirect costs and risks. A hedged position requires less capital, and capital has an interest expense. For example, suppose a derivative has a payoff space of {+19, -21}, with 0.5 probability for an expected value of -1 to the seller. A seller will need at least 21 units of capital. Hedging this asset can reduce this payoff volatility to a certain -1, in which case the seller needs only 1 unit of capital. This illustrates how hedging reduces risk and capital requirements for hedgers but not the average cost of this hedge.

Uniswap LP Profitability

I will start first with a basic derivation of IL for a constant product AMM. I will apply this to the ETH-USD pair, but these formulas apply to any token pair.

We start with the formulas for ETH and USD pool tokens. Liquidity is both an intuitive and technical term within the AMM, and p here would apply the price of ETH in terms of USD.

$$USD_{pool} = liquidity \cdot \sqrt{p_t}$$

$$ETH_{pool} = \frac{liquidity}{\sqrt{p_t}}$$

The value of this LP pool position at any time is given as follows.

$$Value(LP) = USD_{pool} + ETH_{pool} \cdot p_t$$

Substituting for the token amounts given the above pool equations, we get can derive this simply in terms of the price and liquidity for an LP.

$$V\left(LP \mid p_t, liquidity\right) = liquidity \cdot \sqrt{p_t} + \frac{liquidity}{\sqrt{p_t}} \cdot p_t = 2 \cdot liquidity \cdot \sqrt{p_t}$$

Taking the derivative of the LP's pool value with respect to the price, we get the delta

$$delta = \Delta = \frac{\partial V}{\partial p} = \frac{liquidity}{\sqrt{p_t}}$$

The derivative of the delta, or second derivative of the pool value, is thus

$$\frac{\partial^2 V}{\partial p^2} = \frac{\partial \Delta}{\partial p} = gamma = \Gamma = -\frac{liquidity}{2 \cdot p^{3/2}}$$

Given this gamma in the AMM pool, we can apply this to the Black-Scholes formula for convexity cost (gamma/2 times variance), which gives us

$$ConvexityCost = \frac{\Gamma}{2} \cdot p_t^2 \cdot \sigma^2 = \frac{liquidity \cdot \sqrt{p_t} \cdot \sigma^2}{4}$$

The convexity cost equals the option premium via an equilibrium argument where profits are zero: if they were positive, it would not be an equilibrium because sellers would enter the market; if profits were negative, sellers would exit. This argument is used in the famous Black-Scholes equation.

This method of estimating the convexity cost is to determine the option value of an option.

$$Convexity\ Cost = option\ premium$$

This does not mean a convexity payout equals its cumulative convexity cost in every case. A single option payout represents one observation like a draw from a normal distribution reflects the distribution.

$$ATM\ \ straddle\ payoff\ \ = N(optionPremium, \sigma_{optionPayout})$$

For an AMM LP position, which is like an ATM straddle, the convexity cost is the expected IL instead of an actual IL. This is also the IL a good market maker *should* approximate, in that good LPs hedge their delta risk, and hedging does not change the expected IL, but minimizing its variance reduces capital costs.

$$Convexity\ Cost\ =\ E(IL)$$

With the convexity cost function for an AMM, we can apply the daily volume and liquidity data from Uniswap's pools and the actual daily volatility for the assets to calculate the average daily profitability for the LPs of these pools. I used daily liquidity and volume data from two of the most prominent Uniswap pools. I calculated the daily variance using the daily price and minute-downsampled price data daily. I then present the monthly average daily data to see if it's trending. This is in the table below.

Uniswap ETH-USDC Pool Profitability

Monthly data contain average daily values. Average daily volatility was taken from down-sampled minute data. Gross Margin is (revenue-convexity cost)/revenue. Liquidity and volume data are from Uniswap's Ethereum mainnet. Data through March 21.

	daily		0.3% Pool			0.05% Pool		
	volatility	price	volume	liquidity	grossMarg	volume	liquidity	grossMarg
202107	4.5%	2,113	102,043	16,055	-27%	339,403	10,589	-55%
202108	4.3%	3,099	158,558	22,086	-22%	428,440	12,265	-51%
202109	5.2%	3,338	129,893	11,990	-18%	454,525	8,102	-37%
202110	4.0%	3,820	108,169	16,262	-27%	476,796	12,518	-31%
202111	3.7%	4,444	92,952	15,472	-18%	681,058	23,932	-60%
202112	4.1%	4,048	144,264	20,346	-37%	860,335	25,996	-74%
202201	4.6%	3,054	111,027	11,377	13%	822,799	24,070	-57%
202202	4.5%	2,869	91,026	11,836	-20%	808,409	22,931	-61%
202203	3.5%	2,877	67,735	14,221	-13%	637,949	27,195	-35%
202204	3.1%	3,170	65,969	16,585	-16%	625,817	27,260	-20%
202205	5.0%	2,185	119,423	12,745	-8%	817,706	20,101	-22%
202206	6.3%	1,390	121,279	10,350	1%	596,718	11,653	-15%
202207	5.4%	1,358	83,713	11,267	-19%	581,017	15,301	-41%
202208	4.4%	1,699	63,243	10,884	-17%	579,512	18,213	-22%
202209	4.6%	1,484	72,103	12,552	-22%	422,743	14,226	-37%
202210	2.9%	1,367	35,415	15,123	-19%	359,467	23,462	-10%
202211	5.2%	1,296	73,024	11,587	-52%	603,722	18,567	-56%
202212	2.5%	1,237	14,701	12,668	-107%	223,774	27,207	-55%
202301	3.0%	1,456	27,876	10,137	-8%	375,912	31,343	-45%
202302	2.8%	1,624	26,807	10,844	-11%	432,315	30,993	-18%
202303	4.4%	1,586	40,739	8,765	-64%	730,904	27,614	-68%
average	4.2%	2,358	83,331	13,483	-24%	564,729	20,645	-41%

This chart shows that LPs have consistently lost more money via their IL than they made in trading fees. Worse, there is no trend.

The best explanation for the persistence of LPs despite losing money is that they are not calculating the option cost. This would explain the lack of growth, as smart money is not entering this new market.

Why are LPs so Stupid?

One reason the LP convexity cost is not appreciated is that it is not a direct cash charge. Instead, it's cost relative to a pair of assets, as opposed to a simple USD value, which is uncommon. Consider the LP's pool value, which can be represented as a linear function of the square root of the ETH price.

$$ValueLP(p_t) = 2 \cdot liquidity \cdot \sqrt{p_t}$$

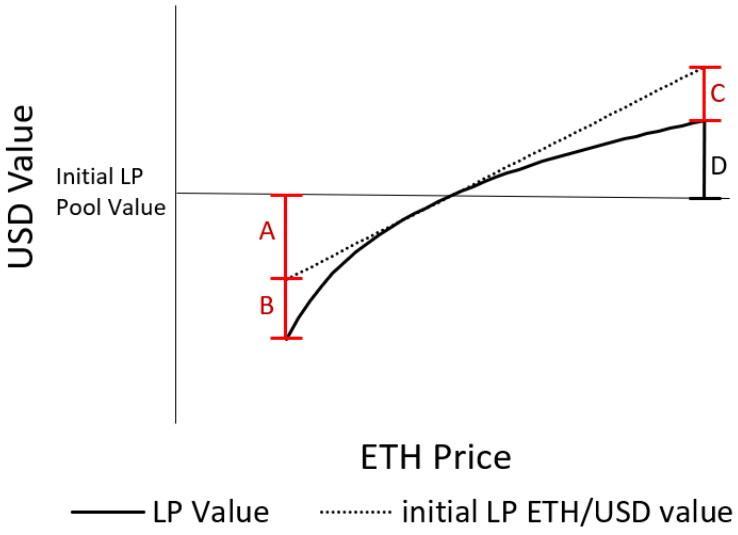
In comparison, the value of their initial deposit is a linear function of price, given an initial deposit of x USD and y ETH.

$$Initial\ LP\ Deposit\ Value\ in\ USD = x + y \cdot p_0$$

The value of the initial deposit and the LP function are equal initially. However, as the LP value function is concave, and the initial portfolio value is linear, we know the LP's future pool value will always be less than the initial portfolio pairs. The value of the LP position as a function of the ETH price is an increasing concave line tangent to the value of the LP's initial deposit, as seen in the figure below. In equilibrium, fees should compensate for this predictable loss.

The daily IL is imperceptible without proper accounting. For example, the daily ETH price volatility is around 4%, and half of an LP's pool value is from ETH, while the IL averages around 0.03%. Further, the benchmark is ambiguous. When the price of ETH rises, the LP's pool value also rises (segment D in the figure below), just less than it would have if not in the pool (segments C+D). When ETH prices decline, the pool position declines (A+B), but so would their original portfolio (A), though less so. To appreciate the option they are implicitly selling, they would have to look at their position relative to an initial position that most only remember at its USD value.

IL Cost Superficially Ambiguous



LPs ignorance is not remedied by more academic studies of LP profitability. Almost all focus on the realized IL of actual LPs assuming none of them hedged. This is like testing option returns by looking at the returns on options independent of the hedge, which no institutional option market maker does (I used to work for one). For example, a significant study by Topaze Blue (Loesch et al., 2021) found that Uniswap pools generated $199.3M in fees over a period that incurred $260M in IL, and 49.5% of LPs lost money. Such a takeaway obscures the profound fact that the LPs lose money because it is tempting to think that those LPs with clever tactics were among the half that made money, and all one has to do is figure out what those tactics are.

Realized ILs will equal the convexity cost over many years, but the realized IL will have much greater volatility in small samples. This is related to why option sellers hedge their positions: reduced volatility reduces risk, which reduces required capital. If option market makers hedge their portfolios, those researching option expenses should use estimates as if the option was hedged. That is implicit in comparisons of implied to future volatility.

Why Realized IL is a Bad Metric

To help see the relative efficiency of these two approaches for estimating IL, I estimated the small-sample properties of both approaches using a Monte Carlo simulation. As I present about 600 daily Uniswap pool observations in my Uniswap LP profitability table above, I generated the price paths over 20 periods of 30 days to get a sense of the sample volatility. I recorded the mean and volatility of the two ways of measuring IL: realized IL based on starting and ending price, and convexity cost based on daily volatility.

Monthly Realized IL in Monte Carlo

$$IL = \left\{USD_{30} + ETH_{30} \cdot p_{30}\right\} - \left\{USD_0 + ETH_0 \cdot p_{30}\right\}$$

I assumed fixed liquidity and had an initial price of 100. As I am interested in comparing one approach to the other, the specific numbers are irrelevant as long as they are the same for both approaches. We are looking at the relative differences generated by these IL estimation methods. The monthly realized IL in the samples can be simplified to the following (they are mathematically identical).

$$IL = -liquidity \cdot \frac{\left(\sqrt{p_{30}} - \sqrt{p_0}\right)^2}{p_0}$$

Monthly Convexity Cost

For the convexity cost approach, I used the basic formula derived above, that is:

$$ConvexityCost = \frac{liquidity \cdot \sqrt{p} \cdot \sigma^2}{4}$$

Applying this using time series of price and volatility generated the following formula for each monthly estimate.

$$ConvexityCost = \frac{liquidity \cdot \sum_{t=0}^{30} \sqrt{p_t} \cdot \sigma_t^2}{4}$$

These two formulas were applied to one million simulations to estimate means and standard

Monte Carlo Estimation of Uniswap LP Costs

ETH minute-down-sampled data from July 2021-March 2023 were used to estimate daily volatility. These daily volatility estimates generated a heteroskedastic random price time series. 20 sets of 30-day realized IL and convexity costs. Both approaches assumed a liquidity=1000 and an initial price of 100.

	ImpLoss	convCost
Mean	1,547	1,604
Stdev	2,302	794

The absolute numbers do not matter, but the relative ones do. The results above show these approaches have approximately equal mean estimates for the IL, as expected. However, the sample IL approach had a **three times larger** standard deviation. Intuitively this makes sense because the convexity cost approach uses daily prices to estimate the next day's price variance, information any hedger would use when managing their convex positions. In contrast, the realized IL approach uses only the start and end prices. As with many options results, many ways exist to prove and intuit these findings. results. The bottom line is that the convexity cost formula dominates the sample IL approach to estimating expected IL (something a good hedge can lock in).

There should not be much doubt that LPs are consistently losing money. Those LPs fortunate not to lose money in the TopazeBlue study were simply random draws that were below average instead of clever.

I don't want to pick on TopazeBlue. Still, if one of the most prominent studies of ILs is misleading, it is understandable most LPs, who are not quants, will not see that LPs lose money outside of random shocks that obscure this loss when not hedged. The lack of LP profitability also explains why well-capitalized groups are not adding liquidity to these AMM pools.

Impermanent Loss Equals Arbitrage Profit

Impermanent loss is the loss due to adverse selection in an LP's pool position, where the pool loses the token with a relative increase in value and gains the token with a relative decline in value. We capture this by comparing the LP's token portfolio composition to its initial token portfolio quantities, both valued at the new price.

$$IL = \{USD_1 + ETH_1 \cdot p_1\} - \{USD_0 + ETH_0 \cdot p_1\}$$

We can substitute for these terms using the equations that define an AMM to generate a more primitive representation as a function of liquidity, starting and ending price, p_0 and p_1.

$$IL = \{USD_1 - USD_0\} + p_1\{ETH_1 - ETH_0\}$$

$$IL = liq \cdot \left(\sqrt{p_1} - \sqrt{p_0}\right) + p_1 \cdot liq \cdot \left(1/\sqrt{p_1} - 1/\sqrt{p_0} \right)$$

$$IL = liq \cdot \left(\sqrt{p_1} - \sqrt{p_0}\right) + p_1 \cdot liq \cdot \left(1/\sqrt{p_1} - 1/\sqrt{p_0} \right)$$

$$IL = liq \cdot \left(2\sqrt{p_1} - \sqrt{p_0} - p_1/\sqrt{p_0} \right)$$

This simplifies to

$$IL = -liq \cdot \frac{\left(\sqrt{p_1} - \sqrt{p_0}\right)^2}{\sqrt{p_0}}$$

The profit generated by an arbitrageur is calculated using the amount of the asset traded; here, the change in ETH from the trader's perspective times the difference between the ending price and the arbitrageur's trade price.

$$ArbitragePnL = \Delta ETH \cdot \left(p_1 - fillPrice \right)$$

$$ArbitragePnL = liq \cdot \left(1/\sqrt{p_1} - 1/\sqrt{p_0} \right) \cdot \left(p_1 - \sqrt{p_1 p_0} \right)$$

Which, via algebra, generates the negative of the IL listed above.

$$ArbitragePnL = liq \cdot \frac{\left(\sqrt{p_1} - \sqrt{p_0}\right)^2}{\sqrt{p_0}}$$

This is one way to intuit that the IL is an actual cost. There are several other ways to see this, such as the note above in LP leverage and liquidation that finds leveraged LP position is generated by its IL.

Financing Rate Fraud

The funding rate mechanism used to link perp prices with spot prices is a farce in that it does not and cannot tie a synthetic price with a spot price via arbitrage, and in practice, it is used to defraud its users. As a practical matter, the perp price is a Schelling point in that its obvious target is the spot price, and the funding rate is just there to make traders feel comfortable that it is not merely a Schelling point. The fact that there is a vague relation to an equilibrating mechanism is good enough for most traders, as they are happy to use centralized platforms like BitMex and DyDx. As in those cases, many users are happy to have access to perps as long as it seems fair.

Given the development of stablecoins, one can trade ETH for USD directly, which allows arbitrage and makes this funding rate mechanism an anachronism. With a leveraged AMM, one

can let natural arbitrage set the price: when the LAMM price is too high, people will sell ETH to the pool to get USDC; when the price is too low, they will sell USDC and get ETH.

One can forgive the perp funding rate scam as its foundational white lie facilitated a much-needed market. In 2016, shorting or leveraging bitcoin was impossible on standard exchanges. All one could do was swap one token on the hybrid exchange EtherDelta. There were no stablecoins or wrapped Bitcoin. BitMex, a centralized unregulated exchange that only took bitcoin deposits, created the first popular perpetual swap, aka 'perp.'

Instead of an expiration date and settlement in a perp market, a perp anchors its price to the spot via a funding rate mechanism. When the perpetual contract's price exceeds the spot price, the story is that this implies longer than short demand. The long traders pay short traders a fee proportional to this price premium to equilibrate the market. Crypto funding rates prevent continuing divergence in the price in perp and spot markets.

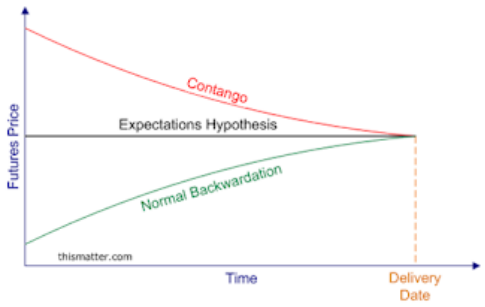
The perp premium is the percent difference between the perp and spot prices. The spot price could be from external markets like Coinbase, or for centralized perps, from spot markets on their exchange:

$$PerpPremium = PerpPrice/SpotPrice - 1$$

The funding rate is like the future expiring once daily, as this premium is applied to 24 hours based on the perp premium. One can apply it to 8-hour windows or anything else, but the standard is to apply the simple premium above and divide it by the number of periods within a day.

For example, suppose you short a BTC perpetual future trading 10% above the underlying index all day. In that case, it's perp premium—then you will receive a total funding payment of 10% over that day to compensate for the fact that, unlike traditional futures markets, there is no expiry or settlement, as it is perpetual. This sounds reasonable, but to understand why this is not, you must first understand the theory of how funding rates work in swap markets or how basis rates work in futures markets.

The basis in futures markets acts as a funding rate in swap markets, defined as the difference between the futures and spot prices. The chart below shows the horizontal time axis moving from a current futures price to its delivery/expiry date if the black line represents the current spot price.



The basis is the difference between the futures and spot price. It can be positive or negative depending on whether the futures price is above or below the spot price. The funding rate is implicit in the amortization of the basis over time, in that, at expiration, the spot price equals the futures price, so the basis is sure to be zero at that time.

There is no basis for swap markets; a funding rate is applied daily, acting precisely like the basis in futures markets. Swap accounts trade at spot prices, facilitated by broker margin. Here the basis goes from being implicit to explicit.

$$LongSwapPnL = Notional\ (p(t+1)/p(t) - 1 - FundingRate)$$

Funding rates in prime broker swap accounts are charged daily and determined independently of the spot prices, like how a bank sets interest rates. For equity swap accounts common among hedge funds, they are generally a fixed markup to the Fed Funds rate, such as adding 25 basis points when a customer borrows USD to go long and subtracting 25 basis points when a customer goes short (which lends USD to the broker).

Thus far, the similarity of swap funding rates and the futures' basis to the perp funding rate seems plausible. Two academic articles are generally referenced when presenting perps. The first is by Gehr (1988), which describes how gold was traded at the Chinese Gold and Silver Exchange Society of Hong Kong (CGSE) in the 1980s. This was when trading was not possible around the clock, and there was no internet, so a price had little volatility outside the trading day. The CGSE was unique because its futures market was undated, i.e., perpetual. The market settled daily and held a 30-minute auction to determine the funding rate. Those long gold compare the cost of paying storage and interest on the spot vs. the funding rate; those short gold futures take delivery if they feel the funding rate is too low. This funding rate was added to the spot price to create a new closing price used in the subsequent day's pnl.

The effect on prices and cash flows in the CGSE futures market was as follows. If the market price closed at 100.00, and the funding rate was determined to be 0.01% over the next day, the cost basis for the next day's PNL is 100.01. If the price stayed constant at 100.00 each day, the longs would lose 0.01 because the daily pnl would be 100.00 – 100.01 on a long position, where 100.00 is the spot close, and 100.01 is the previous day's futures close. The traded price would never be 100.01; it would just be used in the daily calculation of the trader's pnl on the next trading day.

Nobel laureate Robert Shiller (1993) proposed a perpetual futures contract for single-family homes. Unlike a stock index or a commodity, the underlying asset, housing, is challenging to create into a futures commodity because it is not homogeneous like a commodity. Quality varies considerably by location and structure, creating a lemons problem. Shiller proposed a rental index to create a rental return proxy for a housing price index. He proposed a statistical model that correlated with real estate's average rental return, net of depreciation. This rental return would then be paid by the short to the long.

$$s(t+1) = f(t+1) - f(t) + d(t+1) - r \times f(t)$$

In the equation above, *s* is the daily margin change in a trader's account; *f* is the perpetual futures price, *r* is an interest rate adjustment, and *d* is determined outside the market. While this is interesting, the difficulty in generating a robust rental index for *d* is probably why this has never been implemented. The market was supposed to trade at a spot price that did not reflect the daily funding charge, *d*, only its present discounted value. However, rental income, like macroeconomic profit, is challenging to estimate via macroeconomic indicators, and most macroeconomic models work poorly out-of-sample, generating considerable uncertainty for potential traders.

Nonetheless, the estimation method implied that this funding rate would move slowly, like interest rates. There was never the suggestion that the market price reflects the spot price and the funding rate, as there is no spot price in this hypothetical, never-realized market.

The perp-premium charge is *like* Gehr's funding charge, added to the market's spot price after trading. It is also *like* the *d(t+1)* term in Shiller's model. Thus, at 30k feet, the connection between the perp premium and the funding rate seems consistent. However, the average synthetic/spot price ratio is not determining the funding rate in either of these mechanisms, as it is for perps.

In crypto perps, the modal daily funding rate and perp premium is 0.03%, which annualizes to 11%. This is a significant funding rate compared to interest rates that have been near zero over the period where perps have existed. A 0.10% perp premium would imply a massive 36.5% funding rate paid by longs to shorts. The average transaction costs for the most liquid US equities, which are more efficient than any crypto market, are estimated at around 0.1%. This is consistent with Gemini tic data that show a 0.15% standard deviation in the price change from one trade to the next (reflecting a bid-ask bounce).

The perp premium incenting trades at any instant is below the transaction cost, given not just the fees but gas and the effective bid-ask spread, which is paid twice over a round trip. If one were frequently trading, as the price-setting arbitrageurs tend to do, extreme funding rates would be less than a round-trip in transaction costs. For example, a 50% funding rate would imply a 0.006% funding payment for a one-hour position, considerably less than their transaction costs.

Additionally, the perp premium applied to longs and shorts is based on the average perp premium in the future. Even if one could know exactly one's perp premium at the time of trade and transaction costs were zero, it would tell the long-term traders little about what it would be in the future. If one targets positions held for a month, the current perp premium at the trade time is irrelevant.

Supposedly, with all the perp premium's economic insignificance for motivating short- and long-term traders, we expect the market to determine the funding rate by inspiring people to buy and sell perps based on current perp/spot premium movements of 0.02%. This is why it is a farce; it is absurd.

This leads to why the perp premium is consistently positive (payments from longs to shorts) and frequently rises to 40% after crypto prices jump, as it did this week. Market makers dominate

price setting, and all perp markets are effectively centralized and run by unidentified and unaccountable coalitions of insiders. They can target 0.03% or 0.13% above the current spot index. No independent auditors regulate an immutable tape of trades with objective time stamps (as the once-perceived compliance-oriented FTX demonstrated). Anything that can be gamed will be gamed, and perp markets can be gamed.

On average, market makers on standard CLOBs have net zero positions on their assets. On perp exchanges, however, the market makers are generally short because it is much easier for these insiders to hedge their short exchange positions with long positions off the exchange [exchanges have different options depending on the nature and size of other markets on their exchanges, so exceptions exist]. A short hedge would require large amounts of capital on another exchange, generating significant operational risk from regulatory attack surfaces and hackers. This allows the perp market makers a significant extra return on the capital needed for market making.

Below we see that funding rates are positive in bull markets and negative in bear markets. There is an asymmetry in that the positive rates in bull markets are significantly greater than the negative rates in bear markets. Intuitively, they can get away with gouging their longs in bull markets, so they do. In bear markets, they make enough money off their natural short position to afford to give a negative funding rate, as this can be a valuable marketing tool (Move your crypto position here to make an extra 10%!). This allows the perp market makers a significant extra return on the capital needed for market making. While there is a basis on the Bitcoin and ETH CME futures market, it is significantly smaller and probably reflects an arb available for institutional traders with positions on Binance and the CME.

BitMex Annualized Funding Rates in Bull and Bear Markets

	BTC	ETH
'17 bull	22%	n/a
'18 bear	-8%	-19%
6/20-'21 bull	16%	28%
'22bear	-2%	-8%
avg	4%	4%

*ETH funding rate was adjusted for BTC/ETH covariance effect

Theoretically, the perp funding rate should be insignificant, if not zero. Neither USDC nor ETH has a dividend on the blockchain. The cost of carry for USDC and ETH are identical. ETH may have an interest rate if one considers the benefits of staking, but this rate is stable and around 4%, implying a negative funding rate (i.e., longs would get paid to compensate for forgone interest). There are no supply shocks in tokens to generate option value, such as when oil tankers are full. There is nothing like a draught destroying a corn harvest that generates a convenience yield for those with corn inventories. To the extent there is hedging pressure generated by natural long or short ETH, the only natural positions are stakers and miners who are perforce long, implying a negative funding rate (they would pay traders to take their naturally long risk).

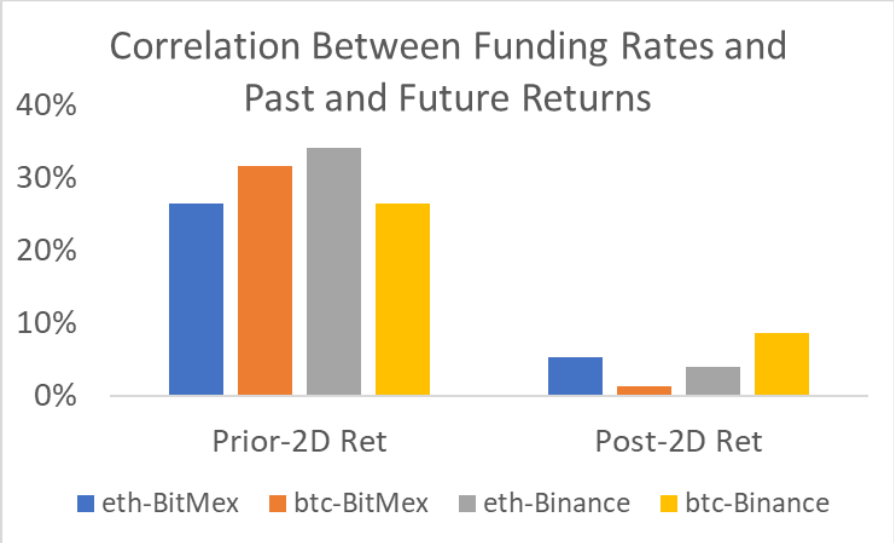
Nothing in the theory of futures basis rates or funding rates implies the large and variable funding rates we observe in perp markets. Standard efficient markets theory, the law of iterated expectations, implies current sentiment is reflected in spot prices, not forward curves. This is why funding rates are generally independent of asset prices, as with equity swap markets or auto loans.

In crypto, perp funding rates are generally strongly positive when the price has risen, as they did the second week of March 2023. This is the opposite of what occurs in commodity markets, where commodity price spikes correspond to negative funding rates, and price declines correspond to positive funding rates. The spot rate moves more than the futures rate in standard futures markets, while the pattern is reversed in perp markets. A unique market mechanism is generating the funding premium in perps.

The chart below shows strong predictability between price changes and future funding rates. The strong positive correlation for prior returns implies that when prices rise, funding rates tend to rise the following day. The contemporaneous and future correlations are almost zero. Funding rates are based on recent returns.

BitMex and Binance FundingRate Correlations

Perps for USD on BitMex, USDT on Binance. BitMex data from 3/'17, BitMex ETH from 9/'18, Binance ETH from 12/'19, Binance BTC from 9/'19.



Crypto perp funding rates are best explained by insider manipulation. When prices generate windfall profits to long perp traders, they do not mind 50% annualized funding rates the following day, which amounts to a mere 0.14% daily charge. It's like how big winners in Vegas often give the dealer a big tip: the *house money* effect. The 50% funding rate premium on perps relative to the regulated and more transparent CME in February 2021 reflects insiders taking what they can from abused customers. Market makers, generally short, receive the funding rate windfall; the game is rigged as heads-they-win-big, tails-they-win-a-little.

The theory that explains the positive return/funding-rate correlation is a typical story that is clear, simple, and wrong. It makes no sense when you get into the details. Like the explanation that price increases come from ‘more buyers than sellers,’ the idea that long demand shows up in futures price premiums has never made sense.

The perp funding rate reflects insider manipulation of customers, a crypto-crypto cost that, if eliminated, would create a superior exchange for those wanting leverage.

Equity Token Rights and Responsibilities

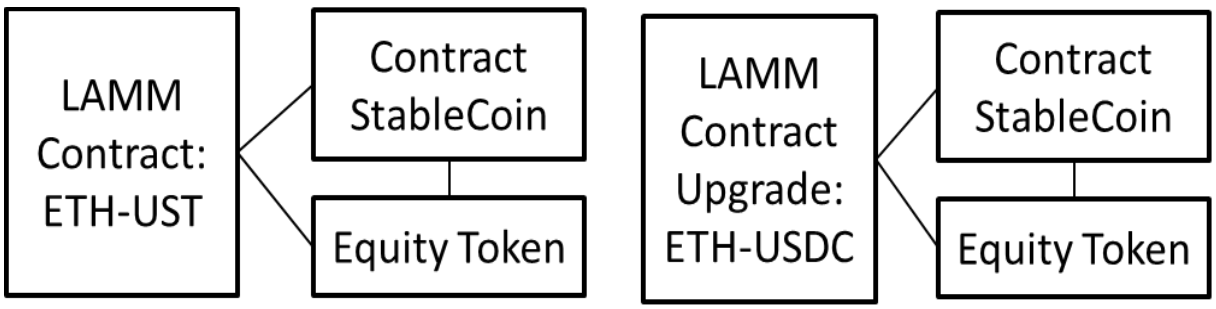
The equity balances in the account have two primary purposes. First, they provide an insurance fund for potential account insolvencies. Second, as the equity account generates revenue from liquidations and trades this aligns the incentives of the equity token holders with the users.

To keep the contract safe, it needs to be simple. Thus, the only way the equity token holders access the contract’s equity is via redemption, which extinguishes the equity tokens and generates a pro-rata distribution of the equity account’s token holdings. For example, a 10% equity token redemption would get 10% of the equity accounts tokens.

If the market price of tokens was above the NAV of the equity account, an equity token holder would be better off swapping their equity token for some other token off the contract. If the NAV of the equity account became too high, so that the expected return on the tokens was too low, this would incent redemptions until the expected return increased sufficiently to give it an equilibrium return. If the NAV was above the traded price of the equity token, redemption would arbitrage this price discrepancy.

Suppose the NAV of the equity account becomes negative. This could be a disaster that removes all future trust in the contract. However, if an outsider thought the contract was still viable, and that it was an anomaly, we want to provide a mechanism for salvaging the contract. In that case, an outsider can recapitalize the contract and then become a large (e.g., 1/4) owner of the equity account, reflecting the value generated by removing this existential risk promptly.

The equity token holders can propose and vote on a new trading contract. The proposal would list the new trading contract’s blockchain address so that people can evaluate its source code. It would also require a minimum amount of equity tokens bonded to prevent frivolous or mischievous proposals. An unsuccessful vote would destroy some or all the bonded equity tokens to prevent frivolous proposals.



A trading contract upgrade could occur for various reasons. For example, owners can change the base stablecoin from USDC to UST, preventing obsolescence caused by changes in the stablecoin market. It could also change merely to add a new coding innovation. In either case, it would proceed via the same voting method on the equity token contract. Only one such upgrade can be evaluated at a time, giving equity token holders the time and focus on evaluating and anticipating this action.

A successful vote on a contract upgrade would then move the internal stablecoin minting rights to the new trading contract, incentivizing existing traders and LPs to move their balances to this new exchange contract. Users could then redeem their stablecoins from the old trading contract, withdraw the USDC, and take their business to the new stablecoins if desired.

As the contract's endogenous stablecoin has only one minter at any time, the stablecoins are not at risk from multiple contracts, each with its own risks. While this limits the upside value of the equity token, and the size of the stablecoin, the safety generated is essential. If the same stablecoin or equity token underlay several contracts, users would have to monitor all of them for potential hacks, and some might be conspicuously riskier than others.