High-Frequency Trading using a Hedging Desk

Izaac Mammadov*

8th November, 2023
(Version 0.1)

Abstract

This article presents an overview of how an internal hedging desk can calculate fair prices to charge for internally holding and then offloading short-term positions in risky assets. It then describes how market-neutral making or taking strategies can use these hedging desk prices to inform what trades it should do. Throughout, it uses the concept of utility functions to quantify the cost of volatility.

1 Utility functions

1.1 Definition

Different people have different ideas about how much variance they are willing to accept in their future wealth for higher expected returns. This is known as their subjective risk-aversion. For example, if forced to choose between two real random variables X_1 or X_2 to set their total wealth equal to, different people, even behaving completely rationally, may choose different variables. More concretely, if X_1 is a constant equal to 100, and $X_2 \sim \mathcal{N}(110, 20^2)$, a more risk-averse person would prefer X_1, while a less risk-averse person would prefer X_2.

Risk-aversion seems on the surface very difficult to quantify in a way that covers all the choices humans may make between any two arbitrary distributions representing their wealth. Thankfully in their 1947 seminal work, von Neumann and Morgenstern essentially solved this problem. They proved the von Neumann-Morgenstern utility theorem, which states that assuming four very weak "rationality" assumptions, then every agent (i.e. person, entity or thing making risk-aversion judgements) has a utility function $U(x)$. Its choices are always based on maximising $\mathbb{E}(U(X))$ over the different X it can choose from.[2] In other words, if an agent is behaving rationally and consistently, then it must have some utility function $U(x)$ and every choice it makes between different distributions for its wealth must just be it choosing the distribution X which maximises $\mathbb{E}(U(X))$.

*Email: izaacmammadov@tuta.io, Website: https://izaac.mammadov.co.uk.

We can further stipulate that in a choice between two constants, agents will always choose the larger constant for their wealth. Hence if $x \geq y$, then $U(x) = \mathbb{E}(U(x)) \geq \mathbb{E}(U(y)) = U(y)$. In other words, U must be an increasing function. Secondly, we can stipulate that for the same expected value, agents prefer certainty over uncertainty, and so they will always prefer their wealth to be a constant $pa + (1-p)b$, compared to having a probability p of being a and a probability $1-p$ of being b. Thus $U(pa+(1-p)b) \geq pU(a)+(1-p)U(b)$, which by definition means U must be concave. Assuming U is twice differentiable, we can simplify these stipulations to just $U'(x) \geq 0$ and $U''(x) \leq 0$ for all x.

1.2 Common utility functions

Someone with no risk-aversion at all will simply always maximise their expected value regardless of variance. If expected values are equal, then they will have no preference between lower or higher variance. This is represented by the utility function $U(x) \equiv x$. Actually, it can be represented by any utility function of the form $U(x) \equiv Ax + B$ with $A > 0$ — these are all equivalent. In general, utility functions do not change behaviour under addition by a constant and multiplication by a positive constant, since $\mathbb{E}(AU(X) + B) = A\mathbb{E}(U(X)) + B$; maximising $\mathbb{E}(AU(X) + B))$ is the same as maximising $\mathbb{E}(U(X))$.

The concavity of the utility function at any point represents in some sense how risk-averse someone is at that level of wealth, and in the case of $U(x) \equiv x$, indeed $U''(x) \equiv 0$. The Arrrow-Pratt absolute risk-aversion coefficient, meant to encapsulate this idea of risk-aversion at a certain point, is defined as

$$A_{\text{abs}}(x) := \frac{-U''(x)}{U'(x)}. \tag{1}$$

[1][3]. Note that the Arrow-Pratt absolute risk-aversion coefficient is invariant under addition by a constant and multiplication by a positive constant. There is also the Arrow-Pratt relative risk-aversion coefficient defined as:

$$A_{\text{rel}}(x) := \frac{-xU''(x)}{U'(x)}, \tag{2}$$

which also has the feature that it is invariant under multiplicaton and addition of the utility function by a constant, but now is also dimensionless to the units of x (it doesn't make a difference if wealth is measured in Dollars, Pounds or Euros).

If we assume that a utility function has constant absolute risk-aversion (CARA), then we are forced to have utility functions of the form $U(x) \equiv -\mathrm{e}^{-cx}$ up to equivalence for some $c > 0$.[1] If we assume that a utility function has constant relative risk-aversion (CRRA), then we are forced to have utility functions of the form $U(x) \equiv \pm x^C$ for some $C \leq 1$,[2] or $U(x) \equiv \log(x)$ up to equivalence.

[1] Not inclding $A_{\text{abs}}(x) \equiv 0$, where $U(x) \equiv x$ up to equivalence

[2] The $\pm$ sign is there to make the function increasing.

1.3 Application to High-Frequency Trading

A high-frequency trading firm will have some concept of wealth (which can be defined to be its total net assets) and will be making expected-value/variance trade-offs every second. Von Neumann-Morgenstern states that (if it is being run rationally and consistently), it must be maximising some expected utility on its total net assets. A common misconception is that trading firms have no risk-aversion and are only trying to maximise expected value ($U(x) \equiv x$), however that would mean that the company should take infinitely sized positive expected-value bets and never hedge. Unfortunately, without an infinite balance sheet, this can only end badly.[3]

What that utility function specifically should be can be a decision left to the stakeholders of the company and all of the discussion below is applicable to all choices of utility function. However, out of all the commonly used utility functions, one stands out as being particularly ubiquitous. That is the CRRA function $U(x) \equiv \log(x)$. It is the only CARA/CRRA function with the properties of:

$$\lim_{x \to 0} U(x) = -\infty, \text{ and } \lim_{x \to \infty} U(x) = \infty. \tag{3}$$

This means that a net assets of 0 should be treated as infinitely bad (indeed, typically meaning the company has very little future left), and that the utility is unbounded in the positive direction, meaning there is always significantly more expected utility to be gained.

However, utility functions were originally conceptualised to help think about trades taken over a fixed time period with risky assets being marked to market at the end of the time period. This is clearly not a good assumption in high-frequency trading, where trades are taken at any time, whenever desired. Thus more care needs to be taken in how utility functions are used which will be decsribed later on.

2 Hedging Desk Behaviour

2.1 Definition

We describe a high-frequency trading set-up as follows. There is a central internal hedging desk which, for every risky asset traded and for all volumes, quotes an internal bid/ask. Every single high-frequency making/taking strategy must immediately (market) trade against the internal hedging desk whenever it receives a net exposure to any risky asset. A trade's profit can be precisely defined by the net cash flow in executing a trade including fees and immediately market trading any exposures against the internal hedging desk. As can be expected, the goal of a making/taking strategy is to maximise the sum of profits across all the trades it executes.

The internal hedging desk, on the other hand, doesn't try to maximise its own profits. Instead, its purpose is to provide as tight as spreads as possible to

[3] cf. Alameda Research.

all of the making/taking strategies while ensuring that it charges enough spread to compensate for the variance in price of the risky assets that it holds. The internal hedging desk doesn't actually execute any external trades but merely quotes internal prices in such a way as to encourage making/taking strategies to make offsetting trades where appropriate.

2.2 Uncertainty from the strategy's perspective

The making/taking strategy faces risk with any trade it executes that the internal hedging desk price may change in the time it takes to receive notice from the exchange of the trade going through and passing on the exposure to the heding desk. For a maker trade, this would be the time between the last point a limit order could've been cancelled for a trade not to go through, and the time the hedging desk receives the trade. For a taker trade, this would simply be the time between sending the order and the time the hedging desk receives the trade.

For the sake of simplicity, let us consider a risky asset XYZ, whose mid-price follows a geometric Brownian Motion, and so, its mid-price $M(t)$ is given by the equation

$$M(t) = M_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma B_t\right), \tag{4}$$

where B_t is a Brownian Motion. We can further assume that we are working in a risk-neutral probability space and so $\mu = 0$, giving us

$$M(t) = M_0 \exp\left(-\frac{\sigma^2 t}{2} + \sigma B_t\right). \tag{5}$$

Then, if we say that our internal hedging desk bid for a volume V of XYZ at time t is $B(V, t)$. We decompose this bid as follows:

$$B(V, t) = M(t)\mathrm{e}^{-C(V,t)}, \tag{6}$$

where $C(V, t)$ is the credits of the hedging desk bid for volume V and time t. If a taker strategy sends a fill or kill buy for XYZ at time t for a volume V, price π, and then the hedging desk receives a confirmation of the trade at time t', the strategy's profit P can be written as follows:

$$\begin{aligned} P &= V \times (B(V, t') - \pi) && (7)\\ &= V \times \left(M(t')\mathrm{e}^{-C(V,t')} - \pi\right) && (8)\\ &= V \times \left(M(t')\mathrm{e}^{-C(V,t')} - M(t)\mathrm{e}^{-C(V,t)} + M(t)\mathrm{e}^{-C(V,t)} - \pi\right) && (9)\\ &= V \times \left(M(t')\mathrm{e}^{-C(V,t')} - M(t)\mathrm{e}^{-C(V,t)}\right) + P_0, && (10) \end{aligned}$$

where P_0 is the profit that the strategy sees/predicts at the time t (when the taker trade is sent). For the sake of simplicity, we assume that $C(V, t') = C(V, t)$, which is a fair assumption if the internal hedging desk doesn't suddenly have significant changes in `XYZ` exposure or predictions around market volatility in the timeframe between t and t'. A more detailed analysis can be done by considering changes in $C(V,t)$ over this short timeframe. Using this assumption, the profit P becomes

$$
\begin{aligned}
P &= V\mathrm{e}^{-C(V,t)}(M(t') - M(t)) + P_0 && (11)\\
&= V\mathrm{e}^{-C(V,t)}M(t)\left(\exp\left(-\frac{\sigma^2(t'-t)}{2} + \sigma(B_{t'} - B_t)\right) - 1\right) + P_0 && (12)\\
&= VB(V,t)\left(\exp\left(-\frac{\sigma^2(t'-t)}{2} + \sigma\sqrt{t'-t}Z\right) - 1\right) + P_0, && (13)
\end{aligned}
$$

where Z is a standard normal distribution. Thus:

$$\frac{P - P_0}{VB(V,t)} + 1 \sim \text{Log-Normal}\left(-\frac{\sigma^2(t'-t)}{2}, \sigma^2(t'-t)\right).^4 \tag{14}$$

This precisely defines the distribution of profits expected based on the size of the trade V, the initial seen hedging desk bid $B(V,t)$, the volatility of the underlying asset σ and the time taken from sending the trade to the hedging desk receiving confirmation $t' - t$. We can thus calculate that the following statistics for P at time t:

$$\mathbb{E}(P) = P_0, \tag{15}$$

$$\mathrm{Var}(P) = V^2B(V,t)^2\left(\mathrm{e}^{\sigma^2(t'-t)} - 1\right). \tag{16}$$

We use this information to figure out if a given trade is worth doing or not based on a specified utility function and current net-assets. Say that a trading firm's current net-assets is given by X_0. Then the trade is worth doing, if at time t, $\mathbb{E}(U(X_0 + P)) \geq \mathbb{E}(U(X_0))$. For a log utility function, that corresponds to:

$$\mathbb{E}(\log(X_0 + P)) \geq \log(X_0) \tag{17}$$

$$\log(X_0) + \mathbb{E}\left(\log\left(1 + \frac{P}{X_0}\right)\right) \geq \log(X_0) \tag{18}$$

$$\mathbb{E}\left(\log\left(1 + \frac{P}{X_0}\right)\right) \geq 0. \tag{19}$$

(20)

This can be calculated via numerical integration for the P as specified in Equation 14, to yield a result of whether a trade is worth doing based on X_0, σ, $t' - t$, P_0, $VB(V,t)$. In general, we can see that as X_0 grows, we are more willing to do riskier trades so long as the expected value is positive.

[4] Using the notation Log-Normal(μ, σ^2).

Note that in this subsection, we have made some key assumptions. These include, the mid-price being a geometric Brownian Motion with no drift; constant hedging desk credits between t and t; only considering Fill or Kill orders; knowing $t' - t$ before the trade takes place; and most incorrectly, the lack of any adverse selection. We will work on dropping or improving these assumptions later on.

2.3 Uncertainty from the hedging desk's perspective

A hedging desk faces uncertainty from having exposure to risky assets. The hedging desk only locks in profit after buying a risky asset and then selling the risky asset or vice versa. In the time between those trades, it faces exposure to the risky asset. The longer the time is between the position-increasing trade, and the position-decreasing trade, the greater the risk it faces. We consider the hedging desk to be a first-in, first-out (FIFO) queue, where each position-decreasing trade hedges against the oldest position-increasing trade still in the queue.

Let us say that our hedging desk currently has a position in `XYZ` of $+A$ (long) and trades the asset `XYZ` at a rate $2r$.[5] If we assume a symmetrical flow of buys and sells, then hedging trades of `XYZ` take place at a rate of r. If the hedging desks further buys an extra V amount of `XYZ`, then that new trade will take an average time of approximately $\frac{2A+V}{2r}$ to hedge. That is calculated from the fact that it must hedge all of its initial $+A$ position (the hedging desk is a FIFO queue) which takes time $\frac{A}{r}$, and then it begins hedging the V amount of `XYZ` which completes after $\frac{V}{r}$. Note that the average time to hedge the trade isn't the same as the time taken to hedge the entire amount V, but rather the time taken to hedge half of the amount V. Because it is a FIFO queue, it doesn't matter if more buy trades take place before hedging has finished, since those trades will just go behind this one.

We work out how much credit to charge for hedging this trade V based on the distribution of $M(t + \frac{2A+V}{2r})$ given by:

$$M\left(t + \frac{2A+V}{2r}\right) = M(t) \exp\left(-\frac{\sigma^2(2A+V)}{4r} + \sigma\sqrt{\frac{2A+V}{2r}}Z\right), \tag{21}$$

where Z is a standard normal variable. We define the profit from the trade to be as follows:

$$P = V\left(M\left(t + \frac{2A+V}{2r}\right) - \exp(-C(V,t))M(t)\right) \tag{22}$$

$$P = VM(t)\left(\exp\left(-\frac{\sigma^2(2A+V)}{4r} + \sigma\sqrt{\frac{2A+V}{2r}}Z\right) - \exp(-C(V,t))\right), \tag{23}$$

[5] Trade rate can have units of dollars per hour, for example.

from which we can calculate the following statistics for the profit of a hedging desk trade:

$$\mathbb{E}(P) = VM(t)(1 - \exp(-C(V,t))) \tag{24}$$

$$\mathrm{Var}(P) = V^2 M(t)^2 \left(\exp\left(\frac{\sigma^2(2A+V)}{2r} \right) - 1 \right). \tag{25}$$

As expected, the expected profit increases with larger credits, and the variance in profit increases with larger initial position A and volume of trade V, while the variance decreases with larger hedge rate r. As before, we can see whether such a trade is worth doing based on our utility function framework if $\mathbb{E}(U(X_0+P)) \geq \mathbb{E}(U(X_0))$, which for a log utility function corresponds to:

$$\mathbb{E}(\log(X_0 + P)) \geq \log(X_0) \tag{26}$$

$$\mathbb{E}\left(\log\left(1 + \frac{P}{X_0} \right) \right) \geq 0, \tag{27}$$

which can once again be calculated via numerical integration. However, in the case of a hedging desk, our goal is to quote prices as tight as possible without losing expected utility, in order to best facilitate the strategies' trading. Thus, in this case, $C(V,t)$ should be chosen to force equality in Equation 27.

In this subsection so far, we have only described the statistics around a position-increasing trade, however, what should we make of a trade that reduces our risk? Once again, let us assume that our current position is $+A$ (long) and we make a trade this time to sell V ($V \leq A$) units of `XYZ` at a price of $M(t)\mathrm{e}^{C(V,t)}$. The credits $C(V,t)$ should be chosen so that our expected utility is equal before and after the trade. This leads to the equation:

$$\mathbb{E}\left(\log\left(X_0 + AM\left(t + \frac{A}{2r} \right) \right) \right) = \mathbb{E}\left(\log\left(X_0 + VM(t)\mathrm{e}^C + (A-V)M\left(t + \frac{A-V}{2r} \right) \right) \right) \tag{28}$$

$$\mathbb{E}\left(\log\left(1 + \frac{A}{X_0} M\left(t + \frac{A}{2r} \right) \right) \right) = \mathbb{E}\left(\log\left(1 + \frac{V\mathrm{e}^C}{X_0} M(t) + \frac{A-V}{X_0} M\left(t + \frac{A-V}{2r} \right) \right) \right), \tag{29}$$

where X_0 is the net assets not including `XYZ`, and

$$M(t + \Delta t) = M(t) \exp\left(-\frac{\sigma^2 \Delta t}{2} + \sigma \Delta t Z \right), \tag{30}$$

with Z once again a standard normal random variable.

Note that our expected utility logic naturally discounts the value of risky assets held by the hedging desk. This creates natural "skewing" behaviour whereby if the hedging desk is significantly long a risky assset, its bids will have large credits, while its offers will have small or negative credit. And since strategies make trades based on the prices quoted by the hedging desk, this will lead to strategies naturally skewing their bids and offers based on the heding desk's exposure. No extraneous skewing logic needs to be implemented if your trading is based on proper risk-discounted valuations for risky-assets, as we have described above.

Once again, we need to point out important assumptions in the model described in this subsection. These are the fact that the mid-price is a geometric Brownian Motion with no drift; hedging takes place at a constant rate r; treating the entire trade volume as being hedged at its "median" hedge time; and once again, most incorrectly, no adverse selection.

3 Implementing Expected Utility

3.1 Taylor Approximations

Above, we've made passing references to using numerical integration to work out values for expected utilities. While that is the only way (for most utility functions) to come to a highly accurate result, in a high-frequency trading scenario, a much quicker although imprecise method is necessary. The solution to this is to use Taylor's Theorem to approximate our utility functions as small order polynomials. The smaller the size of the trade in question, the more accurate these approximations will be, and the fewer terms you'll need to still get an accurate result.

Specifically focusing on a log utlity function, we often see an expression of the form

$$\mathbb{E}(\log(X_0 + Y)) = \log(X_0) + \mathbb{E}\left(\log\left(1 + \frac{Y}{X_0}\right)\right), \tag{31}$$

where Y is some random variable involving a trade and X_0 is a constant representing our net assets. The second term on the right hand side of Equation 31 can be Taylor expanded as follows:

$$\mathbb{E}\left(\log\left(1 + \frac{Y}{X_0}\right)\right) \approx \frac{1}{X_0}\mathbb{E}(Y) + \mathcal{O}\left(\frac{Y^2}{X_0^2}\right) \tag{32}$$

$$\approx \frac{1}{X_0}\mathbb{E}(Y) - \frac{1}{2X_0^2}\mathbb{E}\left(Y^2\right) + \mathcal{O}\left(\frac{Y^3}{X_0^3}\right) \tag{33}$$

$$= \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{iX_0^i}\mathbb{E}\left(Y^i\right), \tag{34}$$

with convergence if $|Y| < |X_0|$ almost surely.[6]

The first-order approximation in Equation 32 involves only considering the expected value of Y. In general, this is the justifcation for why many trading-firms see themselves as solely in the business of maximising expected value — that is what they are doing, though only to a first order approximation. Another way to think about this, is that if you zoom in close enough to any (differentiable) utility function at any point, it will just look like a straight line — no risk-aversion and solely maximising expected value. The second-order approximation in Equation 33 has the same expected-value term, but now

[6]The proof of this statement follows from Fubini's Theorem. The details are left as an exercise to the reader.

subtracts that by a variance term. Indeed, the larger the variance is relative to X_0, the greater this risk-discounting effect is. In general, each term in the Taylor expansion just involve higher and higher moments of Y, which are usually analytically known.

3.2 Defining "net assets"

Throughout, we've made reference to this value X_0 and defined it as the net assets of the trading firm at any point in time. In our context, that is actually an ambiguous definition, since at any point in time, a company may have exposures to a wide variety of risky assets. The entire basis of our system involves not marking risky assets to market, but rather to mark them to their "expected utility" at the predicted time that they get hedged. The net assets value should represent the expected utility of assets if they are diligently passively liquidated, with no time-pressure, which certainly isn't mark-to-market.

Based on this rationale, we propose the following definition for X_0:

$$X_0 := \exp\left(\mathbb{E}\left(\log\left(\sum_{a\in R} A_a M_a\left(t_a\right)\right)\right)\right), \tag{35}$$

where R is the set of all assets, A_a is the exposure to asset a, t_a is the expected hedge time for the current position in a, and $M_a(\tau)$ is the mid-price of asset a at time τ. Using the assumptions about hedging rates as in Subsection 2.3, we see this can be written as:

$$X_0(t) = \exp\left(\mathbb{E}\left(\log\left(\sum_{a\in R} A_a M_a\left(t + \frac{A_a}{2r_a}\right)\right)\right)\right), \tag{36}$$

where r_a is the hedging rate of asset a. Further applying the drift-free Brownian Motion assumption, this turns the equations into:

$$X_0(t) = \exp\left(\mathbb{E}\left(\log\left(\sum_{a\in R} A_a M_a(t)\exp\left(-\frac{A_a\sigma_a^2}{4r_a} + \sigma_a\sqrt{\frac{A_a}{2r_a}}Z\right)\right)\right)\right), \tag{37}$$

where σ_a is the volatility of asset a and Z is a standard normal random variable.

Note that if our exposure was only to one or more non-risky assets (e.g. `USD` if that is what profit is measured in), then we return to the conventional definition of net assets being just the sum of all asset exposures. That is the reason that we need to take the exponential at the start of the expression. We also note that in all of the discussion above, the conditions for whether trades were worth making or not are precisely equivalent to whether the trade increases this specific definition of X_0.

4 Assumption Busting

4.1 Brownian Motion

A very natural assumption to bust is the idea that the Brownian Motion for the mid-price of an asset has constant volatility σ. Instead, we can say that the volatility is function of time and so

$$dM_t = \sigma(t)dB_t, \tag{38}$$

where M is the mid-price and B is a standard Brownian Motion (with volatility 1). This $\sigma(t)$ can be fit in wide variety of ways. For example, we may try to fit the volatility to periodic functions with periods over a day (to model how volatility changes at different times of the day) and over a week (to model how volatility changes on different different days of the week). The volatility can also be based on very recent low-latency observations of market-wide and asset-specific volatility.

4.2 Constants r and $t' - t$

The rate at which hedging trades take place clearly isn't constant and depend on wide variety of factors, including the periodic factors mentioned for volatility. This can be modelled and fit based on historical trading. Another factor that will affect rate of hedging r is the amount of inventory of a certain token available. Once we start running low on inventory, r should rapidly tend towards 0.

The time between sending an order and the hedging desk hearing a confirmation has a very large variance. However, again this can be modelled based on historical data as a log-normal variable. Then for every expression involving $t' - t$, we can simply take the expected value averaging over the distribution of $t' - t$.

4.3 No adverse selection

In order to model adverse selection, we have to build on top of the model given in Equation 38, by adding a negative drift term and an increased variance term. Thus in a short period of time immediately after a trade on a certain exchange E, takes place, the mid-price follows the stochastic process:

$$dM_t = \mu_E dt + (\sigma(t) + \kappa_E)dB_t, \tag{39}$$

where $\mu_E \leq 0$ and $\kappa_E \geq 0$ can be fitted according to historical observations. In fact, the "short period of time" description can also be fitted so that the μ_E and κ_E decay exponentially to 0 as time after the trade increases, with exponential half-life τ_E.

References

[1] Kenneth Joseph Arrow. *Aspects of the theory of risk-bearing.* Yrjö Jahnssonin Säätiö, 1965.

[2] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition).* Princeton University Press, Dec. 2007. DOI: `10.1515/9781400829460`. URL: `https://doi.org/10.1515/9781400829460`.

[3] John W. Pratt. "Risk Aversion in the Small and in the Large". In: *Econometrica* 32.1/2 (Jan. 1964), p. 122. DOI: `10.2307/1913738`. URL: `https://doi.org/10.2307/1913738`.