

# General Research Notes

Zachary Ross

March 3, 2022

# Contents

<b>I</b>	<b>Dated Findings</b>	<b>2</b>
	February 2, 2022 . . . . .	3
	February 9, 2022 . . . . .	3
	February 23, 2022 . . . . .	3
<b>II</b>	<b>Neural Network Pruning</b>	<b>5</b>
<b>1</b>	<b>Pruning neural networks without any data by iteratively conserving synaptic flow [4]</b>	<b>7</b>
1.1	Pruning Algorithms . . . . .	7
1.2	Synaptic Saliency . . . . .	8
1.3	Algorithm . . . . .	8
1.4	Commentary . . . . .	9
<b>III</b>	<b>Gradients</b>	<b>10</b>
<b>2</b>	<b>Efficient Per-Example Gradient Computations [2]</b>	<b>11</b>
2.1	Commentary . . . . .	12
<b>3</b>	<b>Understanding gradient clipping in incremental gradient methods [3]</b>	<b>13</b>
<b>IV</b>	<b>Finance</b>	<b>17</b>
<b>4</b>	<b>An Introduction to Quantitative Finance [1]</b>	<b>18</b>
4.1	Interest Rates . . . . .	18

Part I

Dated Findings

## February 2, 2022

In the work presented to me, Dr. Lee primarily uses pruning for gradient masking rather than parameter masking, although most studies I've read seem to have done the opposite. I initially assumed that given the standard linear layer function with masking

$$\mathbf{z} = (\mathbf{M} \odot \mathbf{W})\mathbf{h} + \mathbf{b} \quad (1)$$

the effect of the mask would still result in non-zero values in the gradient of the matrix. This is falsified by the derivation

$$\frac{\partial L}{\partial W_{i,j}} = \frac{\partial L}{\partial z_i} \frac{\partial z_i}{\partial W_{i,j}} = M_{ij} \frac{\partial L}{\partial z_i} h_j \quad (2)$$

or in vectorized form we have that

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{M} \odot \left( \frac{\partial L}{\partial \mathbf{z}} (\mathbf{h})^\top \right) \quad (3)$$

which implies that the use of a forward mask guarantees that of a backward. The method used instead in Dr. Lee's examples was  $\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}$  with masking only applied during the gradient (i.e. that of Equation 3). So what needs to be tested is whether or not we can apply masking solely during the forward phase i.e. implement Equation 1 with builds its gradient without the mask.

## February 9, 2022

Thorough testing has disproved our previous hypothesis. Strangely enough, the gradient masking is much more effective than weight masking for some odd reason.

## February 23, 2022

In the following discussion, I use beats per minute (BPM) to refer to heart rate. I only do it this way because I've already written it out and don't feel like changing it.

I've been pondering a bit recently on a hypothesis I've had, that being the perception of time is relative to the individual based on several factors. At first I believed that it's connected to BPM, but this was disproved by my addiction to coffee. I've noticed this relativity even in my day to day, some days seem faster than others while certain activities seem to be slower. For example, under the influence of coffee I feel that my day goes by much faster, which leads me to believe the effects of caffeine are misunderstood (which I will go into later). After or during an intense workout, my perception of time seems to slow way the hell down. I notice this because I'll be listening to music, and maybe it has to do with my percussionist background with intense focus on steady timing, but the songs pass by much slower. This is incredibly curious.

On other days where I don't quite feel myself, I notice instead that the music tends to be much quicker, often times leading me to feel like I'm missing alot.

For the coffee thing, maybe its affect on BPM does slow down perception of time, but caffeine could effectively and periodically cut out clips of time, making

us believe that time is moving faster. I believe this affects productivity because it allows us to move from task to task seemingly "quicker". I don't believe it fully alters our productivity; rather, it seems like it allows us to reach the reward much quicker and thereby reinforce the thought process that we are "getting shit done". This activity/reward behavior most likely becomes a self-fulfilling cycle as the momentum continues, motivating us to continue doing tasks.

I've also considered how we would test these hypotheses because it's rather elusive. Say we tried to get a set of people to listen to a metronome prior to an energy intensive activity and post said activity and asked them to compare the two for which one is faster. For one, how would we get reliable information? If we were just to pick random people from a population, they may not even be capable of discerning between two tempos. I suppose we could do a preliminary exam where we vet out those who could not tell the difference even between minor changes in tempo. We could pick out trained musicians, but then the study is limited to musicians technically. I guess we would just have to do the vetting procedure and only select the upper echelon of people which can tell extremely minute differences in tempo after a select period of time.

Thoughts I had while working out, lol.

# **Part II**

## **Neural Network Pruning**

In the following research, we make use frequently of the set of parameters  $\theta$ . A couple things to note about how this is used: it is common practice to reference the set of parameter in a multitude of ways, none of which is extrapolated upon in any papers I've come across and has been entirely frustrating to decipher on its own.  $\theta$  is often referred to a set of parameters rather than a collection, yet is often assigned an ordering of arbitrary dimensionality. I will attempt to disclose my understanding of this dimensionality and these short-hands briefly

1.  $\theta_i$  is access to some arbitrary element of  $\theta$ ,
2.  $\theta_i^{(l)}$  is the  $i^{\text{th}}$  row in the matrix of parameters at layer  $l$ ,
3. and  $\theta_{i,j}^{(l)}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of the matrix of parameters at layer  $l$ . This is periodically shorthanded to  $\theta_{i,j}$  when authors assume an arbitrary layer.

Unless otherwise mentioned, these will be used throughout.

# Chapter 1

## Pruning neural networks without any data by iteratively conserving synaptic flow [4]

This paper, while theoretically captivating, is lackluster with providing mathematical commentary on the concepts and proofs. It may be worthwhile to go back and either notate the more mathematical definitions of the paper, or derive my own definitions and reprove their claims using a more solid theoretical foundation. This may even set a better groundwork for future analysis.

### 1.1 Pruning Algorithms

Pruning algorithms are generally defined by

1. scoring parameters by some metric, and
2. masking parameters according to their scores.

The latter is generally done by removing the parameters (e.g. making the value zero via Hadamard product). This can either be done via global masking or layer-masking, with global-masking performing better but suffering from *layer-collapse*, which is when an entire layer is masked.

Let  $\theta$  be a collection of network parameters and  $\theta_{\text{prune}}$  be the parameters remaining after pruning. The *compression ratio*  $\rho$  of a pruning algorithm is

$$\rho = \frac{|\theta|}{|\theta_{\text{prune}}|} \quad (1.1)$$

. We define  $\rho_{\text{max}}$  as the maximal possible compression ratio for a *network* that doesn't lead to layer collapse and the  $\rho_{\text{cr}}$  as the maximal compression ratio a given



*algorithm* can achieve without inducing layer collapse. Note that the distinction in the two is  $\rho_{\max}$  is maximal for a network while  $\rho_{\text{cr}}$  is maximal for an algorithm. These definitions motivate the following axiom:

**Axiom 1** (Maximal Critical Compression). *For any pruning algorithm and any network, we should always have  $\rho_{\text{cr}} = \rho_{\max}$ .*

## 1.2 Synaptic Saliency

*Synaptic saliency* is a class of score metrics defined by

$$\mathcal{S}(\theta) = \frac{\partial \mathcal{R}}{\partial \theta} \odot \theta \quad (1.2)$$

where  $\mathcal{R}$  is a scalar loss function of the output  $y$  of a feed-forward network parameterized by  $\theta$ . This metric satisfies two conservation laws:

**Theorem 1** (Neuron-wise Conservation of Synaptic Saliency). *For a feedforward neural network with continuous, homogeneous activation functions  $\phi(x) = \phi'(x)x$ , let  $j$  be the index of a hidden neuron in layer  $i$ . The sum of the synaptic saliency for the incoming parameters to a hidden neuron  $\left(\mathcal{S}_j^{(l)\text{in}} = \left\langle \frac{\partial \mathcal{R}}{\partial \theta_j^{(l)}}, \theta_j^{(l)} \right\rangle\right)$  is equal to the sum of the synaptic saliency for the outgoing parameters from the hidden neuron  $\left(\mathcal{S}_j^{(l)\text{out}} = \left\langle \frac{\partial \mathcal{R}}{\partial \theta_{:,j}^{(l+1)}}, \theta_{:,j}^{(l+1)} \right\rangle\right)$ .*

**Theorem 2** (Network-wise Conservation of Synaptic Saliency). *The sum of the synaptic saliency across any set of parameters that exactly separates the input neurons  $x$  from the output neurons  $y$  of a feedforward neural network with homogeneous activation functions equals  $\left\langle \frac{\partial \mathcal{R}}{\partial x}, x \right\rangle = \left\langle \frac{\partial \mathcal{R}}{\partial y}, y \right\rangle$ .*

## 1.3 Algorithm

For the following theorem, we define the *prune size* to be the total score for the parameters pruned at any iteration and the *cut size* to be the total score for an entire layer.

**Theorem 3.** *If a pruning algorithm with global-masking assigns positive scores that respect layer-wise conservation, and if the prune size is strictly less than the cut size whenever possible, then the algorithm satisfies the Maximal Critical Compression axiom.*

The pruning algorithm defined in this paper uses a loss function

$$\mathcal{R}_{\text{SF}} = \mathbb{1}^\top \left( \prod_{l=1}^L |\theta^{(l)}| \right) \mathbb{1} \quad (1.3)$$

so that for a fully connected network ( $f_\theta(x) = \theta^{(L)} \dots \theta^{(1)}x$ ) the Synaptic Flow score for a parameter  $\theta_{i,j}^{(l)}$  is

$$\mathcal{S}_{\text{SF}}(\theta_{i,j}^{(l)}) = \left[ \mathbf{1}^\top \prod_{k=l+1}^L |\theta^{(k)}| \right]_i |\theta_{i,j}^{(l)}| \left[ \prod_{k=1}^{l-1} |\theta^{(k)}| \mathbf{1} \right]_j \quad (1.4)$$

i.e. the portion of the  $l_1$ -path norm the network has through this parameter. This contributes to the development of Algorithm 1.

---

**Algorithm 1** Iterative Synaptic Flow Pruning (SynFlow)

---

**Require:** network  $f_\theta$ , compression ratio  $\rho$ , iteration steps  $n$

```

 $\mu = \mathbf{1}$  ▷ Initialize binary mask
for  $k$  in  $[1, \dots, n]$  do
     $\theta_\mu \leftarrow \mu \odot \theta_0$  ▷ Mask parameters
     $\mathcal{R} \leftarrow \mathbf{1}^\top \left( \prod_{l=1}^L |\theta_\mu^{(l)}| \right) \mathbf{1}$  ▷ Evaluate SynFlow objective
     $\mathcal{S} = \frac{\partial \mathcal{R}}{\partial \theta_\mu} \odot \theta_\mu$  ▷ Compute SynFlow score
     $\tau \leftarrow (1 - \rho^{-k/n})$  percentile of  $\mathcal{S}$  ▷ Find threshold
     $\mu \leftarrow (\tau < \mathcal{S})$  ▷ Update mask
end for
return  $f_{\mu \odot \theta}$  ▷ Return masked network

```

---

## 1.4 Commentary

A thought I had: the loss function defined in Equation 1.3 seems to almost indicate a just-past-the-minimum metric for the predefined network. It effectively is just the sum of the entries in the product of the network space. Although I agree with the data agnostic approach, I feel there should be a more suitable loss function which should seek to preserve this score rather than minimize it, or at the very least, should compare this score against some optimum, i.e. the optimum score for that network under those parameters. I also feel that the prune size and cut size should have a greater impact over the outcome of the resulting network, such that the compression ratio is fairly equal for all layers.

**Research Question 1.** Does SynFlow, or other pruning methods for that matter, produce the same network given different initial parameterizations? Furthermore, might we be able to use a genetic algorithm to find the “lottery ticket”?

One thing that needs to be considered with this question is that it’s entirely possible that isomorphic networks could be produced, so there may need to be some kind of check that occurs post pruning that confirms whether or not this network has been discovered already.

# Part III

## Gradients

## Chapter 2

# Efficient Per-Example Gradient Computations [2]

An important note about this paper is that it is *not* a derivation of the per-example gradient itself. Rather, it “describes an efficient technique for computing the *norm* of the gradient” with respect to the loss function.

Let each layer of a neural network be defined by the standard transformations

$$\begin{aligned}\mathbf{z}^{(i)} &= \mathbf{W}^{(i)} \mathbf{h}^{(i-1)} \\ \mathbf{h}^{(i)} &= \phi^{(i)} \left( \mathbf{z}^{(i)} \right)\end{aligned}\tag{2.1}$$

with the bias assumed to be an extra row/column of  $\mathbf{W}$  and loss function  $\mathcal{L}$ . Let  $\mathcal{B} \subset \mathcal{D}$  be a minibatch of the input dataset  $\mathcal{D}$ , and let  $\mathcal{B}^{(j)}$  be the  $j$ th example in the minibatch with  $\mathcal{L}^{(j)}$  corresponding to the loss of  $\mathcal{B}^{(j)}$  and cost function  $C = \sum_{j=1}^n \mathcal{L}^{(j)}$ . The per example gradient norm is computed with respect both to the example and the layer, such that the gradient norm for layer  $i$  and example  $j$  is

$$\left\| \frac{d\mathcal{L}^{(j)}}{d\mathbf{W}^{(i)}} \right\|^2 = \sum_{k,l} \left( \frac{\partial \mathcal{L}^{(j)}}{\partial W_{k,l}^{(i)}} \right)^2, \tag{2.2}$$

which is just the frobenius norm. Note that this equation can be used to calculate the  $L_2$  norm of the parameter gradient for an example or the  $L_2$  norm of the gradient for an individual weight matrix by summing over  $j$  or  $i$ , respectively.

The proposed method makes use of  $\mathbf{H}^{(i)}$  whose  $j$ th row contains the activation layer  $\mathbf{h}^{(i)}$  corresponding to  $\mathcal{B}^{(j)}$  i.e.  $\mathbf{H}_j^{(i)} = \phi^{(i)} \left( \mathbf{W}^{(i)} \phi^{(i-1)} \left( \dots \left( \mathbf{W}^{(1)} \mathcal{B}^{(j)} \right) \dots \right) \right)$ . Likewise, define  $\mathbf{Z}^{(i)}$  for each  $\mathbf{z}^{(i)}$  and compute  $\bar{\mathbf{Z}} = \nabla_{\mathbf{Z}} C$ , which can be computed in a single pass using standard backpropagation. This gives us

$$\left\| \frac{\partial \mathcal{L}^{(j)}}{\partial \mathbf{W}^{(i)}} \right\|^2 = \left( \sum_k \left( \bar{Z}_{j,k}^{(i)} \right)^2 \right) \left( \sum_k \left( H_{j,k}^{(i-1)} \right)^2 \right) = \left\| \bar{\mathbf{Z}}_j^{(i)} \right\|^2 \left\| \mathbf{H}_j^{(i-1)} \right\|^2. \tag{2.3}$$

## 2.1 Commentary

This method works well with gradient clipping due to the efficient computation of gradient norms. This calculation allows for simple re-scaling of the gradients used in backwards propagation.

## Chapter 3

# Understanding gradient clipping in incremental gradient methods [3]

This paper considers the class of problems dealing with the minimization of

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{where} \quad f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x), \quad (3.1)$$

where  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  are differentiable, non-convex functions. As a side note, this differs from notation that I'm used to, in that  $x$  is considered to be the set of parameters while  $i$ , or more importantly  $f_i$ , is a function of the parameters using the  $i^{\text{th}}$  component of the dataset to measure the loss. That is, the function  $f_i(x)$  determines the loss accumulated when using  $x$  as the set of parameters on input  $i$ .  $f$  is said to be the *objective function* while each  $f_i$  is called a *component function*.

The incremental method for SGD is

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) \quad k = 0, 1, \dots \quad (3.2)$$

where  $x_k$  is the values of the parameters at iteration  $k$ ,  $\alpha_k$  is the learning rate, and  $f_{i_k}$  is a uniform randomly chosen component function  $i_k \in \{1, 2, \dots, m\}$  of the dataset. The incremental method for IGC is

$$x_{k,i} = x_{k,i-1} - \alpha_k \nabla f_i(x_{k,i-1}) \quad i = 1, \dots, m \quad k = 0, 1, \dots \quad (3.3)$$

where the recursion is defined by base case  $x_{0,0}$  and iterative step  $x_{k,0} = x_{k-1,m}$  and is more cyclical by nature.

The clipping function  $\mathcal{C} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with clipping threshold parameter  $\eta > 0$  is defined by

$$\mathcal{C}(g; \eta) = \min \left\{ 1, \frac{\eta}{\|g\|} \right\} \cdot g. \quad (3.4)$$

Our iterative methods can take advantage of clipping by clipping the gradient so that SGD with clipping is

$$x_{k+1} = x_k - \alpha_k \mathcal{C}(\nabla f_{i_k}(x_k); \eta), \quad (3.5)$$

and IGC with clipping is

$$x_{k,i} = x_{k,i-1} - \alpha_k \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad (3.6)$$

and is written iteratively (using the notation  $x_k = x_{k,0} = x_{k-1,m}$ ) as

$$x_{k+1} = x_k - \alpha_k \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad (3.7)$$

For the theoretical understanding of the clipping function's effects on the convergence of both SGD and IGC, we make the following assumptions:

**Bounded below** there exists an  $f^*$  such that for all  $x \in \mathbb{R}^n$ ,  $f(x) \geq f^*$ ;

**Relaxed smoothness** there exist constants  $L_0, L_1 \in \mathbb{R}^+$  such that  $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$ .

The second condition becomes Lipschitz smoothness when  $L_1 = 0$ . Let  $f_k = f(x_k)$  and assume for simplicity from here on out that SGD and IGC have a constant step size  $\alpha$  and that IGC's notation for  $x_{k,i}$  can be simplified for  $x_{k,0}$  to  $x_k$ . The paper shows that convergence for the algorithm can be ensured if there exists a constant  $C$  such that

$$\left\langle \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k \right\rangle \geq C \|\nabla f_k\|^2 \quad (3.8)$$

i.e. if the magnitude of the clipped gradient projected in the direction of the regular gradient is bounded below by the magnitude of the regular gradient. This value is simplified further as follows.

Let  $g_{ki} = \nabla f_i(x_k)$  and note the following properties of  $g_{ki}$

$$g_{ki} = \beta_{ki} \nabla f_k + t_{ki} \quad \text{where} \quad \beta_{ki} = \frac{\langle g_{ki}, \nabla f_k \rangle}{\|\nabla f_k\|^2} = \frac{\|g_{ki}\|}{\|\nabla f_k\|} \cos \theta_{ki} \quad (3.9)$$

$$t_{ki} \in \nabla f_k^\perp = \{z \mid \langle z, \nabla f_k \rangle = 0\}$$

Since  $\frac{1}{m} \sum_i g_{ki} = \nabla f_k$ , it is a necessary condition that  $\sum_i \beta_{ki} = m$  and  $\sum_i t_{ki} = 0$ . Let  $\gamma_{ki} = \min\{1, \eta/\|g_{ki}\|\}$  and note that  $\mathcal{C}(g_{ki}; \eta) = \gamma_{ki} g_{ki}$ . The paper then shows that

$$\left\langle \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k \right\rangle = \sum_{i=1}^m \gamma_{ki} \beta_{ki} \|\nabla f_k\|^2. \quad (3.10)$$

so that Equation 3.8 becomes

$$\sum_{i=1}^m \gamma_{ki} \beta_{ki} \geq C, \quad (3.11)$$

that is, *the necessary condition for convergence is that the sum in Equation 3.11 is greater than some positive scalar  $C$ .*

There are then essentially two conditions, either of which will ensure the convergence of our algorithm:

1. all  $g_{ki}$  can be partitioned into pairs so that if  $g_{kj}$  and  $g_{kl}$  are a pair then  $\beta_{kj} + \beta_{kl} > 0$  and  $\|t_{kj}\| = \|t_{kl}\|$ ;
2. all  $\beta_{ki} > 0$  or equivalently all  $\cos \theta_{ki} > 0$ .

Under the second condition, the paper then shows that a lower bound can be derived making use of the minimum and maximum gradient magnitudes w.r.t. an example  $i$ . That is, if we let  $m_g = \min_i \|g_{ki}\|$  and  $M_g = \max_i \|g_{ki}\|$ , and we let constants  $c_1$  and  $c_2$  be defined by the inequalities

$$0 < c_1 \leq \frac{1}{m} \sum_{i=1}^m \cos \theta_{ki} \quad \text{and} \quad 0 < c_2 \leq \frac{m_g}{M_g}, \quad (3.12)$$

then the lower bounds for Equation 3.8 are then divided into three scenarios:

1. if  $\eta \geq M_g$ , then

$$\left\langle \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k \right\rangle = m \|\nabla f_k\|^2; \quad (3.13)$$

2. if  $\eta \leq m_g$ , then

$$\left\langle \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k \right\rangle = \left( \eta \sum_{i=1}^m \cos \theta_{ki} \right) \|\nabla f_k\|; \quad (3.14)$$

3. and if  $m_g < \eta < M_g$ , then

$$\left\langle \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k \right\rangle \geq m \min\{\eta c_1, c_2 \|\nabla f_k\|\} \|\nabla f_k\|. \quad (3.15)$$

The paper concludes by providing the following theorems on the behavior of SGD and IGC with gradient clipping.

**Theorem 4.** *Assuming that we have a lower bound and relaxed smoothness, if*

$$\alpha \leq \frac{c_1}{4\eta L_1} \quad (3.16)$$

*then the iterates  $\{x_k\}$  generated by Equation 3.5 satisfy*

$$\frac{1}{T+1} \sum_{k=0}^T \min\{\eta c_1, c_2 \|\nabla f_k\|\} \|\nabla f_k\| \leq \frac{2(f_0 - f^*)}{(T+1)\alpha} + 5\alpha\eta^2 L_0 + \frac{\alpha\eta^3 L_1 c_1}{c_2}. \quad (3.17)$$



**Theorem 5.** *Assuming that we have a lower bound, relaxed smoothness, and each component function also satisfies relaxed smoothness i.e.*

$$\|\nabla^2 f_i(x)\| \leq L_{0i} + L_{1i} \|\nabla f_i(x)\|, \quad i = 1, \dots, m; \quad (3.18)$$

if

$$\alpha \leq \min \left\{ \frac{c_1}{2m(2\eta L_1 + G)}, \frac{1}{5mL_{0i}}, \frac{1}{8m\eta L_{1i}} \right\}, \quad i = 1, \dots, m; \quad (3.19)$$

then the iterates  $\{x_k\}$  generated by Equation 3.7 satisfy

$$\frac{1}{T+1} \sum_{k=0}^T \min\{\eta c_1, c_2 \|\nabla f_k\|\} \|\nabla f_k\| \leq \frac{2(f_0 - f^*)}{(T+1)\alpha m} + 5\alpha m \eta^2 L_0 + \frac{\alpha m \eta^2 c_1 (2\eta L_1 + G)}{2c_2}, \quad (3.20)$$

where

$$G = 5 \max_{1 \leq i \leq m} L_{0i} + 16\eta \max_{1 \leq i \leq m} L_{1i}. \quad (3.21)$$

# **Part IV**

# **Finance**

## Chapter 4

# An Introduction to Quantitative Finance [1]

### 4.1 Interest Rates

**Definition 1.** A *notional* or *principal* is an initial deposit or value  $N$ .

**Definition 2.** An *interest rate*  $r$  is the rate at which a value is increased according to a specified frequency of time.

Suppose we have an account with interest rate  $r$  and notional  $N$ . The value of the account at time  $T$  is

$$N_T = Ne^{rT} \quad (4.1)$$

Throughout finance, there exist discrete and continuous analogs for most equations due the discrete case being more realistic while the continuous case is far more mathematically efficient. That being said, discrete analogs will be used sparsely and we will primarily depend on continuous equations. The discrete versions can be derived from the continuous versions by noting that

$$e^{rT} = \left(1 + \frac{r_m}{m}\right)^{mT} \quad (4.2)$$

where  $T$  is commonly denoted as the time in years,  $m$  is the frequency the interest is compounded, and  $r_m$  is the equivalent interest rate with the discrete frequency.  $r_m$  can then be recovered from  $r$  using

$$r_m = m \left( e^{\left(\frac{r}{m}\right)} - 1 \right). \quad (4.3)$$

**Definition 3.** A *money market account* is an account compounded continuously at rate  $r$  with notional 1. I.e.  $M_0 = 1$  and  $M_t = e^{rt}$ .

**Definition 4.** A *zero coupon bond* (ZCB) with maturity  $T$  is an asset that pays 1 at time  $T$  (and nothing else). The value of a ZCB at time  $t$  for a continuously compounded rate  $r$  is denoted as

$$Z(t, T) = e^{-r(T-t)} \quad (4.4)$$

Table 4.1: Daycount conventions and accrual factors

Daycount convention	$\alpha$ for 16 December 2011 - 16 March 2012
act/365	91/365
act/act	15/365 + 76/366
act/360	91/360
30/365	1/4

where  $Z(T, T) = 1$  by definition. The values  $Z(t, T)$  with  $0 \leq t \leq T$  are known as *discount factors* or *present values*.

The intuition behind the present values for a ZCB derives from the fact that if we have two portfolios, one a ZCB with maturity  $T$  and another that contains cash which will accumulate interest to be worth 1 at time  $T$ , then both will be worth the same value at each  $t$  leading up to  $T$  and thereafter.

**Definition 5.** An *annuity* a series of fixed cashflows at specified times  $T_i = 1, \dots, n$ . The value at time  $t$  is given by

$$V = C \sum_{i=1}^n Z(t, T_i). \quad (4.5)$$

When we substitute in Equation 4.4, we notice that this is just a geometric series. Because of this, if an annuity is paid once a year for  $M$  years with rate  $r$ , this equates to

$$V = \sum_{i=1}^n \frac{1}{(1+r)^i} = \frac{1}{r} \left( 1 - \frac{1}{(1+r)^M} \right) \quad (4.6)$$

**Definition 6.** The *accrual factor* is (under the discrete analogy) the frequency which interest is compounded, and is denoted as  $\alpha$ .

In Equations 4.2 and 4.3, this is the same as taking  $\alpha = 1/m$ . Due to the year being uneasily divided by simple values of  $\alpha$ , several *daycount conventions* have been made for different settings. We list these in Table 4.1. The daycount conventions are seemingly deceptive in that they hide various details behind terminology. For example, two cases are 30/x and x/360, which respectively denote that each month in the period should be counted as only having 30 days and each month in a year should be counted as only having 30 days. I want to emphasize the weird convention because when we encounter a fraction in the work such as 30/act, this means to say that if we specify a period of three months, then this fraction actually denotes

$$\frac{30 * 3}{m_1 + m_2 + m_3} \quad (4.7)$$

where each  $m_i$  is the number of actual days in the specified month. Like wise, if we have a daycount act/act, this denotes the number specifies the number of actual days in the period divided by the actual days in the year. If we have a leap year, then the denominator contains 366, whereas if February is included in the

daycount, then we may have 28 or 29 days dependent on the year. I hope this helps to explain it better, if not then *oof*.

A few specific cases arise but are only defined in the exercises. I will list those here.

**Definition 7.** *Semi-bond* is the US dollar interest rate which is semi-annual compounding with 30/360 daycount and is denoted as  $\gamma_{SB}$ .

**Definition 8.** *Annual money* is the US dollar interest rate which is annual compounding with act/360 daycount and is denoted as  $\gamma_{AM}$ .

If we know the rate  $r_\alpha$  and accrual factor  $\alpha$  of a quoted interest rate over a specified period of time, and we are interested in the equivalent rate  $r_\beta$  under a different accrual factor  $\beta$  over the same period of time, then this information can be retrieved by making note of the equivalency

$$r_\alpha \alpha = r_\beta \beta. \quad (4.8)$$

This makes conversions simple as demonstrated by the following example.

**Example 1.** If on 16 December 2011 the act/365 rate for three months is 5%, the act/360 rate is

$$r_\beta = r_\alpha \frac{\alpha}{\beta} = 5\% \frac{\text{act}/365}{\text{act}/360} = 5\% \frac{360}{365} = 4.9315\%. \quad (4.9)$$

**Example 2.** The conversion between semi-bonds and annual money is given by

$$\gamma_{SB} \frac{30}{360} = \gamma_{AM} \frac{\text{act}}{360}, \quad (4.10)$$

so that

$$\gamma_{SB} = \gamma_{AM} \frac{\text{act}}{30}, \quad (4.11)$$

and likewise

$$\gamma_{AM} = \gamma_{SB} \frac{30}{\text{act}}. \quad (4.12)$$

**Definition 9.** A *stock* or *share* is an asset giving ownership in a fraction of a company. The price at time  $t$  of a stock is denoted by  $S_t$ . The current known price is called the *spot*.

**Definition 10.** A *fixed rate bond* with coupon  $c$  and notional  $N$  is an asset that pays a coupon  $cN$  at a specified frequency (usually a year) and  $N$  at its maturity date  $T$ . A *floating rate bond* is defined similarly with a variable interest rate.

# Bibliography

- [1] S. Blyth. *An Introduction to Quantitative Finance*. OUP Oxford, 2013. ISBN: 9780199666591. URL: <https://books.google.com/books?id=SXbcAAAAQBAJ>.
- [2] Ian Goodfellow. *Efficient Per-Example Gradient Computations*. 2015.
- [3] Jiang Qian et al. “Understanding gradient clipping in incremental gradient methods”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1504–1512.
- [4] Hidenori Tanaka et al. “Pruning neural networks without any data by iteratively conserving synaptic flow”. In: *CoRR* abs/2006.05467 (2020). arXiv: 2006.05467. URL: <https://arxiv.org/abs/2006.05467>.