

IEEE Working Group P3109 Interim Report on 8-bit Binary Floating-point Formats

Questions and comments via GitHub issues at
<https://github.com/P3109/Public>

Initial release: 18 September 2023

Version 0.5.1: 24 November 2023

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Inc.

Three Park Avenue

New York, New York 10016-5997, USA

All rights reserved.

This document is subject to change. USE AT YOUR OWN RISK! IEEE copyright statements SHALL NOT BE REMOVED from this draft, or modified in any way. Because this is an unapproved draft, this document must not be utilized for conformance / compliance purposes.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 1.1 | Typographical conventions and notation | 3 |
| 2 | Values | 4 |
| 2.1 | Subnormals | 5 |
| 2.2 | Not a number (NaN) | 5 |
| 2.3 | Zero | 6 |
| 2.4 | Infinities | 6 |
| 2.5 | Extremal Values | 7 |
| 3 | Classification operators | 8 |
| 3.1 | Classifications | 8 |
| 4 | Comparison operators | 9 |
| 4.1 | The totalOrder predicate | 9 |
| A | Rationales | 10 |
| A.1 | Mask Values | 10 |
| A.2 | Overflow to Infinity | 10 |
| B | External Formats | 11 |
| C | Value Tables | 11 |
| C.1 | Value Table: P2, $p = 2$, $\text{emax} = 31$ | 12 |
| C.2 | Value Table: P3, $p = 3$, $\text{emax} = 15$ | 13 |
| C.3 | Value Table: P4, $p = 4$, $\text{emax} = 7$ | 14 |
| C.4 | Value Table: P5, $p = 5$, $\text{emax} = 3$ | 15 |
| C.5 | Value Table: P6, $p = 6$, $\text{emax} = 1$ | 16 |
| C.6 | Value Table: P7 (Linear), $p = 7$, $\text{emax} = 0$ | 17 |

1 Introduction

This document represents ongoing discussions and current matters of consensus from IEEE Working Group P3109, “Standard for Arithmetic Formats for Machine Learning”. The Project Authorization Request (PAR) for P3109 defines the scope, need, and stakeholders as follows:

Scope of proposed standard: This standard defines a binary arithmetic and data format for machine learning-optimized domains. It also specifies the default handling of exceptions that occur in this arithmetic. This standard provides a consistent and flexible arithmetic framework optimized for Machine Learning Systems (MLS) in hardware and/or software implementations to minimize the work required to make MLS interoperable with each other, as well as other dependent systems. This standard is aligned with IEEE Std 754-2019 for Floating-Point Arithmetic.

Need for this Work: Machine Learning Systems have different arithmetic requirements from most other domains. Precisions tend to be lower, and accuracy is measured in dimensions other than just numerical (e.g. inference accuracy). Furthermore, machine learning systems are often integrated into mission-critical and safety-critical systems. With no standards specifically addressing these needs, Machine Learning Systems are built with inconsistent expectations and assumptions that hinder the compatibility and reuse of machine learning hardware, software, and training data.

Stakeholders for the Standard: System developers, vendors, and users of machine learning applications across many industries and interests including but not limited to computation, storage, medical, telecommunications, e-commerce, fleet management, automotive, robotics, and security.

The scope of this interim release is interchange formats only. The working group continues to deliberate on the specification of operations.

1.1 Typographical conventions and notation

Bold text describes the decisions and specifications of this document.

Text that is not bold is background material, typically providing rationale and arguments that represent discussions of the working group leading to a decision and specification.

This document specifies 8-bit floating-point interchange formats (binary formats) and associated operations. Binary formats are parameterized by their width, the number of bits spanned in memory (here, 8); and their precision (p), the number of bits spanned by the true significand (this is one more than the bits of the significand that are stored explicitly).

The formats defined herein shall be referred to as “binary8” formats, and further qualified by precision yielding names “binary8pp”.

For example, “binary8p3” is a format with 3 bits of precision; one bit is an implicit leading bit and two bits are explicit.

2 Values

This section describes the set of values that a binary8 format shall represent. The universe of values in existing floating point usage encompasses some finite real numerical values, the non-finite numerical values positive and negative infinity ($-\text{Inf}$, $+\text{Inf}$), the non-numeric not-a-number values (NaN, NaN_1, \dots), and negative zero (-0.0). The value set for each binary8 format specifies the set of values that are available in that format.

Each binary format shall be associated with a unique encoding. An 8-bit binary encoding is a mapping from 8-bit strings to values. Some of these mappings are included in Appendix C.

Values are considered either “special” or “ordinary”. The special values are Zero, $+\text{Inf}$, $-\text{Inf}$, and NaN. The ordinary values consist of the normal and subnormal values. The special values have encodings that are shared by all binary8 formats:

| Special Value | Hexadecimal Encoding | Bit Sequence |
|-------------------------------------|----------------------|--------------|
| Zero | 0x00 | 0000 0000 |
| Positive Infinity ($+\text{Inf}$) | 0x7F | 0111 1111 |
| Negative Infinity ($-\text{Inf}$) | 0xFF | 1111 1111 |
| Not a Number (NaN) | 0x80 | 1000 0000 |

Table 1: Special Value Encodings

These mappings are shared by all binary8 formats.

| Parameter | binary8p{p} | binary8p5 | binary8p4 | binary8p3 | binary16 | binary32 | binary64 |
|--|-----------------------------------|------------------|------------------|------------------|-----------------|-----------------|-----------------|
| Storage width in bits k | 8 | 8 | 8 | 8 | 16 | 32 | 64 |
| Precision in bits p | p | 5 | 4 | 3 | 11 | 24 | 53 |
| Max exponent emax | $2^{k-p-1} - 1$ | 3 | 7 | 15 | 15 | 127 | 1023 |
| Sign bit | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Exponent field width w | $8 - p$ | 3 | 4 | 5 | 5 | 8 | 11 |
| Exponent bias, bias | $\text{emax} + (p > 1)$ | 4 | 8 | 16 | 15 | 127 | 1023 |
| Trailing significand field width in bits t | $p - 1$ | 4 | 3 | 2 | 10 | 23 | 52 |

Table 2: Parameters for binary formats

Format-defining parameters in bold, derived parameters in normal font.

Adapted from Table 3.5 of IEEE Std 754-2019, and extended to include the binary8pp formats. Concepts are explained in detail in this section.

The finite floating-point numbers representable with a binary format is determined by two *format-defining* parameters:

- Precision p , the number of digits in the significand including the implicit leading bit.
- Maximum exponent emax , the exponent of the largest finite value.

IEEE Std 754-2019 includes the radix b and the minimum exponent emin in a list of format-defining parameters, this document excludes both of them for these two reasons:

- This document covers binary (radix 2) formats only, so b is not a format parameter.

- The quantity $emin$ is determined by p and $emax$; it cannot be varied independently, and so cannot be a format-defining parameter.

P3109 formats shall define $emax(p)$ to be $2^{8-p-1} - 1$. In IEEE-754, $emax$ was set for all defined formats to be $2^{w-1} - 1$, where w is the exponent field width in bits. In this report, this convention is formalized: $emax$ is a fixed function of p , written $emax(p)$, with the formula as given above.

This choice of formula yields the following properties:

- the binary8pp value sets are subsets of the IEEE-754 binary16 value set for $p > 2$
- values are distributed close to symmetrically below and above the value 1.

For $p = 8$, the IEEE-754 formula yields $emax = -\frac{1}{2}$, meaning all non-special values are irrational. Rounding the computation upward yields $emax(8) = 0$, with the consequence that the value sets and encodings for binary8p7 and binary8p8 are identical.

The choice of $emax$ for a given format then determines the exponent bias for that format. The bias is chosen so that the exponent of the largest finite value is $emax$. For IEEE-754 formats, the largest finite value corresponds to an exponent field which has all but the zeroth bit set (e.g. 11110 for binary16), because all of the values with all-bits-one exponents (ABOE values) are occupied by non-finite values (Not-a-Numbers or Infinities). Thus, the biased exponent of the largest finite value is $2^w - 2$, from which bias is computed as

$$emax = (2^w - 2) - bias \implies bias = (2^w - 2) - (2^{w-1} - 1) = 2^{w-1} - 2 + 1 = emax$$

For the binary8 formats in this document where $p > 1$, only one of the ABOE values is non-finite ($\pm Infinity$), so the biased exponent of the largest finite value is $2^w - 1$. Hence the bias calculation becomes

$$bias = (2^w - 1) - (2^{w-1} - 1) = 2^{w-1} - 1 + 1 = emax + 1$$

For $p = 1$, there are zero trailing significand bits, so all ABOE values are special, and again $bias = emax$.

2.1 Subnormals

Binary8 value sets shall include subnormals.

IEEE-754 value sets include subnormals. A value with trailing significand field m and exponent e is interpreted as $1.m \times 2^{e-b}$ except when all bits of the exponent bitfield are 0, in which case the value is $0.m \times 2^{e-b}$.

When training models, it is common to represent near-zero values for gradients. Subnormal numbers induce equal quantization steps around zero; this expands the reach of binary8 trainable models. In statistical applications, the subnormal range is useful for uniform and similar distributions; subnormals are uniformly spaced around zero. They also support working with Gaussian-like distributions, where numbers around zero are more probable.

2.2 Not a number (NaN)

Binary8 value sets shall include exactly one NaN, which shall not signal.

Other floating-point formats define several NaN values, denoted (NaN, NaN₁, ...). NaNs are returned from operations with results outside the set of values. For example, DIV(0, 0), or ADD(Inf, -Inf). Multiple NaN encodings are used in other formats to allow different exceptional conditions to be distinguished.

In the context of machine learning systems, uses of NaN include:

- Debugging of code running on accelerator hardware. In AI accelerators, exceptions may be difficult or expensive to convey back to user code, so it is common practice to allow NaN values to propagate through calculations to indicate that an error has occurred.
- Use as a 'notable value' indicator. In some datasets, for example, tabular data, values may be missing. It is useful to use a value outside the normal numeric range to indicate the position of these values. Particularly when memory usage is a concern, as may be expected in applications where 8-bit formats are being considered, the use of a separate "mask" array, or a list of indices, imposes additional memory overhead. In some cases, Inf can be used as a missing value, but given the restricted range of binary8 formats, it is likely that infinity shall be used as a separate indicator of rounding from values outside of the finite range.
- The use of multiple NaN payloads is known in statistical code (e.g. the R system has NaN and N/A), but it is not widely used, and in the context of binary8, multiple NaNs impose either additional hardware complexity (using only a subset of the significand range), or a large reduction in encoding space (e.g. 8 codes for E5, 16 codes for E4, 32 codes for E3).

2.3 Zero

Binary8 formats shall have exactly one zero. This zero value is nonnegative.

The inclusion of negative zero would incur the cost of an additional code point. Given the decision to encode only a single NaN, placing that NaN at the negative zero code point enables the strictly positive and strictly negative number ranges to be symmetric.

A key rationale for including -0 in IEEE-754 was the consistent implementation of branch cuts in the atan2 function [4, 5]. Although the atan function is common in deep learning, it is generally used as an activation function, rather than a trigonometric operation, and the atan2 function is rare, if not unknown, in deep learning applications. Furthermore, it is not expected that this standard shall define either atan or atan2 .

A secondary reason for providing -0 is the hardware simplification offered by its presence in the implementation of sign/magnitude arithmetic. However, the existence of in-market implementations is evidence that the small hardware simplification has not been sufficient to balance the loss of one code point.

It might be considered that the use of integer comparisons in sorting would argue against placing NaN at the negative zero code point. For example, the JAX machine learning framework is known to sort using integer comparison [3]. However, such sorting still requires $O(n)$ preprocessing and postprocessing steps to enable the use of twos-complement integer comparison, and already has special treatment of NaN and -0 , so eliminating -0 and placing NaN in the -0 position imposes negligible additional burden. Sorting using comparison operations, as typically implemented, is undefined in the presence of NaNs. However, existing practice is to sort NaNs using `totalOrder`.

2.4 Infinities

Binary8 formats shall include positive and negative infinities.

This decision causes a reduction in dynamic range (252 values rather than 254), while offering improved numerical robustness in important machine learning use cases.

Two generic classes of such usage are:

- Mask values, for example, in Transformer models in machine learning [1].

- Representation of overflow.

As illustrated in Appendix A, both usages are facilitated by the presence of infinity.

2.5 Extremal Values

| Format | minSubnormal | maxSubnormal | minNormal | maxNormal | maxFinite |
|-----------|--------------------|-----------------------|--------------------|---------------------|---------------------|
| binary8p2 | 1×2^{-32} | 1×2^{-32} | 1×2^{-31} | 1×2^{31} | 1×2^{31} |
| binary8p3 | 1×2^{-17} | $3/2 \times 2^{-16}$ | 1×2^{-15} | $3/2 \times 2^{15}$ | $3/2 \times 2^{15}$ |
| binary8p4 | 1×2^{-10} | $7/4 \times 2^{-8}$ | 1×2^{-7} | $7/4 \times 2^7$ | $7/4 \times 2^7$ |
| binary8p5 | 1×2^{-7} | $15/8 \times 2^{-4}$ | 1×2^{-3} | $15/8 \times 2^3$ | $15/8 \times 2^3$ |
| binary8p6 | 1×2^{-6} | $31/16 \times 2^{-2}$ | 1×2^{-1} | $31/16 \times 2^1$ | $31/16 \times 2^1$ |
| binary8p7 | 1×2^{-6} | $63/32 \times 2^{-1}$ | 1×2^0 | $63/32 \times 2^0$ | $63/32 \times 2^0$ |

Table 3: Extremal values

It is practical to offer extremal finite values for supported 8-bit binary interchange formats. Following IEEE 754-2019 naming patterns, we adopt: $\text{maxNormal}(T)$, $\text{minNormal}(T)$, $\text{minSubnormal}(T)$ where T is a binary8 format. For example: $\text{maxNormal}(\text{binary8p4}) = 7/4 \times 2^7$, $\text{minNormal}(\text{binary8p5}) = 1 \times 2^{-3}$.

Table 3 shows these values in binary8 formats for $1 < p < 8$.

3 Classification operators

Conforming implementations shall provide these classification predicates and the classifier function. The classification predicates and the classifier function shall not signal exceptions.

The classification operators comprise: 1) a set of functions with a boolean return value, taking a single binary8 value as input; 2) a function `class(x)` that returns a single value of enumeration type, describing the input value's properties.

Predicates shall behave as follows:

| Predicate | Definition |
|--------------------------|---|
| <code>isZero</code> | iff ^a <code>x</code> is 0 |
| <code>isNaN</code> | iff <code>x</code> is NaN |
| <code>isInfinite</code> | iff <code>x</code> is infinite |
| <code>isFinite</code> | iff <code>x</code> is zero, subnormal or normal |
| <code>isNormal</code> | iff <code>x</code> is normal, hence finite |
| <code>isSubnormal</code> | iff <code>x</code> is subnormal |
| <code>isSignMinus</code> | iff <code>x</code> has a negative sign ^b |
| <code>isCanonical</code> | True ^c |
| <code>isSignaling</code> | False ^d |

^aiff abbreviates "if and only if"

^b`isSignMinus(NaN)` is True: NaN is 0x80 (0b10000000).

^cThere are no non-canonical binary8 interchange formats.

^dAll binary8 formats have one NaN; it does not signal.

Table 4: Classification Predicates

3.1 Classifications

The Classifier function `class(x)` shall return enumeration values as follows:

| Enumeration | Condition |
|-------------------|--|
| NaN | <code>isNaN(x)</code> |
| Zero | <code>isZero(x)</code> |
| positiveInfinity | <code>isInfinite(x)</code> and not(<code>isSignMinus(x)</code>) |
| positiveNormal | <code>isNormal(x)</code> and not(<code>isSignMinus(x)</code>) |
| positiveSubnormal | <code>isSubnormal(x)</code> and not(<code>isSignMinus(x)</code>) |
| negativeInfinity | <code>isInfinite(x)</code> and <code>isSignMinus(x)</code> |
| negativeNormal | <code>isNormal(x)</code> and <code>isSignMinus(x)</code> |
| negativeSubnormal | <code>isSubnormal(x)</code> and <code>isSignMinus(x)</code> |

Table 5: Classifier Logic

4 Comparison operators

Conforming implementations shall provide the following comparison operators and the `totalOrder(x, y)` function.

Comparison operators are two argument predicates and their negations that return True or False. Comparisons shall not raise exceptions. Comparisons may be ordered or unordered. A comparison is unordered iff either argument is NaN. All other comparisons are ordered.

For $\{=, >, \geq, <, \leq, \leqslant\}$, if any argument is NaN, the result is False.

For $\{\neq, \not>, \not\geq, \not<, \not\leq, \not\leqslant\}$, if any argument is NaN, the result is True.

Otherwise, the result of a comparison shall match the mathematical result.

| math symbol | predicate <i>true relations</i> | math symbol | negation <i>true relations</i> |
|----------------|---|-----------------------------------|--|
| = | CompareEqual <i>equal</i> | \neq , NOT = | CompareNotEqual <i>less than, greater than, unordered</i> |
| > | CompareGreater <i>greater than</i> | $\not>$, NOT > | CompareNotGreater <i>less than, equal, unordered</i> |
| \geq | CompareGreaterEqual <i>equal, greater than</i> | $\not\geq$, NOT \geq | CompareLessUnordered <i>less than, unordered</i> |
| < | CompareLess <i>less than</i> | $\not<$, NOT < | CompareNotLess <i>greater than, equal, unordered</i> |
| \leq | CompareLessEqual <i>less than, equal</i> | $\not\leq$, NOT \leq | CompareGreaterUnordered <i>greater than, unordered</i> |
| \leqslant | CompareOrdered <i>less than, equal, greater than</i> | $\not\leqslant$, NOT \leqslant | CompareUnordered <i>unordered</i> |

Table 6: Comparison Predicates and Negations

4.1 The totalOrder predicate

`totalOrder(x, y)` provides a total ordering over each binary8 format's value set.

The predicate `totalOrder(x, y)` shall return { True, False } in accord with the logic given below. It shall not raise any exceptions.

```
boolean totalOrder(x, y)
    if isNaN(x): return True
    if isNaN(y): return False
    return compareLessEqual(x, y)
end
```

Note: Following 754's definition of `totalOrder(x, y)`, binary8 NaNs (0x80) compare as the most negative value. The most significant bit of NaN is set, so to be consistent with 754, NaN is ordered before all numerical values.

A Rationales

This section is being rewritten

A.1 Mask Values

A common use for ∞ is to create masks, for example, in Transformer models in machine learning [1].

These values, assembled in mask matrix M with values $M_{ij} \in \{0, -\infty\}$ are typically added to computed values A , in a computation such as:

$$\log(\text{sum}(\exp(\tau * (A + M))))$$

where τ is a “temperature” or “base” parameter [2]. This calculation depends on the property $\exp(\tau * (A_{ij} - \infty)) = 0$.

If a floating point encoding does not provide infinity, then instead M_{ij} will be replaced by a large float (e.g. 224 is the largest finite binary8p4 value). This is not in itself a difficulty: if all the A values are bounded (e.g. the results of a softmax operation are bounded above by 1.0), then $\exp(1.0 - 224.0)$ is an extremely small number, which will certainly round to zero. Therefore, an explicit representation of infinity is *not* needed in order for this computation to yield its desired value.

However, careful implementations do not execute the calculation as written, and instead fuse the $\log(\text{sum}(\exp(v)))$ operation into a single operation $\text{logsumexp}(v)$, whose implementation makes use of the identity transformation

$$\text{logsumexp}(v) \rightarrow \text{logsumexp}(v - \max(v)) + \max(v)$$

Without the “sticky” properties of Inf, this would produce incorrect answers. For example, in a format where $\text{maxFinite}=240$ without Inf, and $\text{maxFinite}=224$ with Inf:

$$\text{logsumexp}(\tau * [-224, -\infty]) \rightarrow \text{logsumexp}(\tau * [0, -\infty])$$

while

$$\text{logsumexp}(\tau * [-224, -240]) \rightarrow \text{logsumexp}(\tau * [0, -16])$$

If $\tau = 1$ and all calculations are done in 8-bit floating point, then the answer will be the same, because $\exp(-16) \approx 1.1 \times 10^{-7}$, which will round to zero in all precisions $p > 2$; but if τ is small, or calculations are done in mixed precision, as is common with 8-bit floating point, the loss of “stickiness” will silently yield unexpected answers. It is not expected that the full calculation shall be done in 8-bit floating point, but the subtraction of the maximum value (and computation of the maximum) might reasonably be in 8-bit floating point.

A.2 Overflow to Infinity

A second use of infinity is to indicate overflow on conversion to the binary8 type. Existing implementations offer several behaviors on overflow: overflow to infinity, saturation to MaxFloat, and overflow to NaN. The existence of a code point for infinity allows any of these options to be implemented in a given instantiation, while removing the code point removes the possibility of implementing the first.

B External Formats

This table summarizes the points of difference and agreement between the formats proposed in this document and a number of existing formats, some of which have hardware implementations.

OCP: Open Compute Platform [6], describing hardware implementations including nVidia, Intel, and ARM.

AGQ: AMD, Graphcore, Qualcomm[7], implemented in Graphcore's C600 product, and AMD's gfx940.

TSL: Tesla Dojo Technology [8], A Guide to Tesla's Configurable Floating Point Formats & Arithmetic

| Format | P3109 | | | OCP | | AGQ | | TSL | |
|---|-------|----|----|-----|----|-----|----|-----|-----|
| Subformat | P3 | P4 | P5 | E5 | E4 | E5 | E4 | E4 | E5 |
| Special values shared by all subformats | | Y | | N | | Y | | N | |
| Exactly one NaN | | Y | | N | | Y | | Y | |
| Positive and negative infinity | | Y | | N | Y | N | | N | |
| Include negative zero | | N | | N | | Y | | N | |
| Max exponent emax | 15 | 7 | 3 | 15 | 8 | 15 | 7 | N/A | N/A |

C Value Tables

Value tables mapping 8-bit strings to value sets are provided in this section.

A typical entry is of the form:

HEX = BINARY = BINARY_FLOAT = DECIMAL
 0x0b = 0.0001.011 = +0b1.011 $\times 2^{-7}$ = 0.0107421875

Where the fields are interpreted as follows:

HEX Hexadecimal encoding of the code point
 BINARY Binary expansion of the code point, underscores separate `sign`, `exponent`, `significand`
 BINARY_FLOAT The precise float value as a binary fraction followed by 2^e with decimal exponent `e`
 DECIMAL A decimal expansion of the value. If the expansion is not an exact representation of the precise float value, the equals sign is replaced by “approximately equals” \approx .

In addition, entries for subnormal and special values are rendered in color as follows:

0x05 = 0.0000.101 = +0b0.101 $\times 2^{-7}$ = 0.0048828125 Subnormal value
 0x80 = 1.0000.000 = NaN Special value (NaN, +Inf, -Inf)

C.1 Value Table: P2, $p = 2$, $\text{emax} = 31$

| | | | |
|--|---|---|--|
| 0x00 = 0.000000.0 = 0.0 | 0x40 = 0.100000.0 = +0b1.0×2 ⁰ = 1.0 | 0x80 = 1.000000.0 = NaN | 0xc0 = 1.100000.0 = -0b1.0×2 ⁰ = -1.0 |
| 0x01 = 0.000000.1 = +0b0.1×2 ⁻³¹ ≈ 2.3283064E-10 | 0x41 = 0.100000.1 = +0b1.1×2 ⁰ = 1.5 | 0x81 = 1.000000.1 = -0b0.1×2 ⁻³¹ ≈ -2.3283064E-10 | 0xc1 = 1.100000.1 = -0b1.1×2 ⁰ = -1.5 |
| 0x02 = 0.000001.0 = +0b1.0×2 ⁻³¹ ≈ 4.6566129E-10 | 0x42 = 0.100001.0 = +0b1.0×2 ¹ = 2.0 | 0x82 = 1.000001.0 = -0b1.0×2 ⁻³¹ ≈ -4.6566129E-10 | 0xc2 = 1.100001.0 = -0b1.0×2 ¹ = -2.0 |
| 0x03 = 0.000001.1 = +0b1.1×2 ⁻³¹ ≈ 6.9849193E-10 | 0x43 = 0.100001.1 = +0b1.1×2 ¹ = 3.0 | 0x83 = 1.000001.1 = -0b1.1×2 ⁻³¹ ≈ -6.9849193E-10 | 0xc3 = 1.100001.1 = -0b1.1×2 ¹ = -3.0 |
| 0x04 = 0.000010.0 = +0b1.0×2 ⁻³⁰ ≈ 9.3132257E-10 | 0x44 = 0.100010.0 = +0b1.0×2 ² = 4.0 | 0x84 = 1.000010.0 = -0b1.0×2 ⁻³⁰ ≈ -9.3132257E-10 | 0xc4 = 1.100010.0 = -0b1.0×2 ² = -4.0 |
| 0x05 = 0.000010.1 = +0b1.1×2 ⁻³⁰ ≈ 1.3969839E-09 | 0x45 = 0.100010.1 = +0b1.1×2 ² = 6.0 | 0x85 = 1.000010.1 = -0b1.1×2 ⁻³⁰ ≈ -1.3969839E-09 | 0xc5 = 1.100010.1 = -0b1.1×2 ² = -6.0 |
| 0x06 = 0.000011.0 = +0b1.0×2 ⁻²⁹ ≈ 1.8626451E-09 | 0x46 = 0.100011.0 = +0b1.0×2 ³ = 8.0 | 0x86 = 1.000011.0 = -0b1.0×2 ⁻²⁹ ≈ -1.8626451E-09 | 0xc6 = 1.100011.0 = -0b1.0×2 ³ = -8.0 |
| 0x07 = 0.000011.1 = +0b1.1×2 ⁻²⁹ ≈ 2.7939677E-09 | 0x47 = 0.100011.1 = +0b1.1×2 ³ = 12.0 | 0x87 = 1.000011.1 = -0b1.1×2 ⁻²⁹ ≈ -2.7939677E-09 | 0xc7 = 1.100011.1 = -0b1.1×2 ³ = -12.0 |
| 0x08 = 0.000010.1 = +0b1.0×2 ⁻²⁸ ≈ 3.7252903E-09 | 0x48 = 0.100100.0 = +0b1.0×2 ⁴ = 16.0 | 0x88 = 1.000100.0 = -0b1.0×2 ⁻²⁸ ≈ -3.7252903E-09 | 0xc8 = 1.100100.0 = -0b1.0×2 ⁴ = -16.0 |
| 0x09 = 0.000100.1 = +0b1.1×2 ⁻²⁸ ≈ 5.5879354E-09 | 0x49 = 0.100100.1 = +0b1.1×2 ⁴ = 24.0 | 0x89 = 1.000100.1 = -0b1.1×2 ⁻²⁸ ≈ -5.5879354E-09 | 0xc9 = 1.100100.1 = -0b1.1×2 ⁴ = -24.0 |
| 0x0a = 0.000101.0 = +0b1.0×2 ⁻²⁷ ≈ 7.4505806E-09 | 0x4a = 0.100101.0 = +0b1.0×2 ⁵ = 32.0 | 0x8a = 1.000101.0 = -0b1.0×2 ⁻²⁷ ≈ -7.4505806E-09 | 0xca = 1.100101.0 = -0b1.0×2 ⁵ = -32.0 |
| 0x0b = 0.000101.1 = +0b1.1×2 ⁻²⁷ ≈ 1.1175871E-08 | 0x4b = 0.100101.1 = +0b1.1×2 ⁵ = 48.0 | 0x8b = 1.000101.1 = -0b1.1×2 ⁻²⁷ ≈ -1.1175871E-08 | 0xcb = 1.100101.1 = -0b1.1×2 ⁵ = -48.0 |
| 0x0c = 0.000110.0 = +0b1.0×2 ⁻²⁶ ≈ 1.4901161E-08 | 0x4c = 0.100110.0 = +0b1.0×2 ⁶ = 64.0 | 0x8c = 1.000110.0 = -0b1.0×2 ⁻²⁶ ≈ -1.4901161E-08 | 0xcc = 1.100110.0 = -0b1.0×2 ⁶ = -64.0 |
| 0x0d = 0.000110.1 = +0b1.1×2 ⁻²⁶ ≈ 2.2351742E-08 | 0x4d = 0.100110.1 = +0b1.1×2 ⁶ = 96.0 | 0x8d = 1.000110.1 = -0b1.1×2 ⁻²⁶ ≈ -2.2351742E-08 | 0xcd = 1.100110.1 = -0b1.1×2 ⁶ = -96.0 |
| 0x0e = 0.000111.0 = +0b1.0×2 ⁻²⁵ ≈ 2.9802322E-08 | 0x4e = 0.100111.0 = +0b1.0×2 ⁷ = 128.0 | 0x8e = 1.000111.0 = -0b1.0×2 ⁻²⁵ ≈ -2.9802322E-08 | 0xce = 1.100111.0 = -0b1.0×2 ⁷ = -128.0 |
| 0x0f = 0.000111.1 = +0b1.1×2 ⁻²⁵ ≈ 4.4703484E-08 | 0x4f = 0.100111.1 = +0b1.1×2 ⁷ = 192.0 | 0x8f = 1.000111.1 = -0b1.1×2 ⁻²⁵ ≈ -4.4703484E-08 | 0xcf = 1.100111.1 = -0b1.1×2 ⁷ = -192.0 |
| 0x10 = 0.001000.0 = +0b1.0×2 ⁻²⁴ ≈ 5.9604645E-08 | 0x50 = 0.101000.0 = +0b1.0×2 ⁸ = 256.0 | 0x90 = 1.001000.0 = -0b1.0×2 ⁻²⁴ ≈ -5.9604645E-08 | 0xd0 = 1.101000.0 = -0b1.0×2 ⁸ = -256.0 |
| 0x11 = 0.001000.1 = +0b1.1×2 ⁻²⁴ ≈ 8.9406967E-08 | 0x51 = 0.101000.1 = +0b1.1×2 ⁸ = 384.0 | 0x91 = 1.001000.1 = -0b1.1×2 ⁻²⁴ ≈ -8.9406967E-08 | 0xd1 = 1.101000.1 = -0b1.1×2 ⁸ = -384.0 |
| 0x12 = 0.001001.0 = +0b1.0×2 ⁻²³ ≈ 1.1920929E-07 | 0x52 = 0.101001.0 = +0b1.0×2 ⁹ = 512.0 | 0x92 = 1.001001.0 = -0b1.0×2 ⁻²³ ≈ -1.1920929E-07 | 0xd2 = 1.101001.0 = -0b1.0×2 ⁹ = -512.0 |
| 0x13 = 0.001001.1 = +0b1.1×2 ⁻²³ ≈ 1.7881393E-07 | 0x53 = 0.101001.1 = +0b1.1×2 ⁹ = 768.0 | 0x93 = 1.001001.1 = -0b1.1×2 ⁻²³ ≈ -1.7881393E-07 | 0xd3 = 1.101001.1 = -0b1.1×2 ⁹ = -768.0 |
| 0x14 = 0.001010.0 = +0b1.0×2 ⁻²² ≈ 2.3841858E-07 | 0x54 = 0.101010.0 = +0b1.0×2 ¹⁰ = 1024.0 | 0x94 = 1.001010.0 = -0b1.0×2 ⁻²² ≈ -2.3841858E-07 | 0xd4 = 1.101010.0 = -0b1.0×2 ¹⁰ = -1024.0 |
| 0x15 = 0.001010.1 = +0b1.1×2 ⁻²² ≈ 3.5762787E-07 | 0x55 = 0.101010.1 = +0b1.1×2 ¹⁰ = 1536.0 | 0x95 = 1.001010.1 = -0b1.1×2 ⁻²² ≈ -3.5762787E-07 | 0xd5 = 1.101010.1 = -0b1.1×2 ¹⁰ = -1536.0 |
| 0x16 = 0.001011.0 = +0b1.0×2 ⁻²¹ ≈ 4.7683716E-07 | 0x56 = 0.101011.0 = +0b1.0×2 ¹¹ = 2048.0 | 0x96 = 1.001011.0 = -0b1.0×2 ⁻²¹ ≈ -4.7683716E-07 | 0xd6 = 1.101011.0 = -0b1.0×2 ¹¹ = -2048.0 |
| 0x17 = 0.001011.1 = +0b1.1×2 ⁻²¹ ≈ 7.1525574E-07 | 0x57 = 0.101011.1 = +0b1.1×2 ¹¹ = 3072.0 | 0x97 = 1.001011.1 = -0b1.1×2 ⁻²¹ ≈ -7.1525574E-07 | 0xd7 = 1.101011.1 = -0b1.1×2 ¹¹ = -3072.0 |
| 0x18 = 0.001100.0 = +0b1.0×2 ⁻²⁰ ≈ 9.5367432E-07 | 0x58 = 0.101100.0 = +0b1.0×2 ¹² = 4096.0 | 0x98 = 1.001100.0 = -0b1.0×2 ⁻²⁰ ≈ -9.5367432E-07 | 0xd8 = 1.101100.0 = -0b1.0×2 ¹² = -4096.0 |
| 0x19 = 0.001100.1 = +0b1.1×2 ⁻²⁰ ≈ 1.4305115E-06 | 0x59 = 0.101100.1 = +0b1.1×2 ¹² = 6144.0 | 0x99 = 1.001100.1 = -0b1.1×2 ⁻²⁰ ≈ -1.4305115E-06 | 0xd9 = 1.101100.1 = -0b1.1×2 ¹² = -6144.0 |
| 0x1a = 0.001101.0 = +0b1.0×2 ⁻¹⁹ ≈ 1.9073486E-06 | 0x5a = 0.101101.0 = +0b1.0×2 ¹³ = 8192.0 | 0x9a = 1.001101.0 = -0b1.0×2 ⁻¹⁹ ≈ -1.9073486E-06 | 0xda = 1.101101.0 = -0b1.0×2 ¹³ = -8192.0 |
| 0x1b = 0.001101.1 = +0b1.1×2 ⁻¹⁹ ≈ 2.8610229E-06 | 0x5b = 0.101101.1 = +0b1.1×2 ¹³ = 12288.0 | 0x9b = 1.001101.1 = -0b1.1×2 ⁻¹⁹ ≈ -2.8610229E-06 | 0xdb = 1.101101.1 = -0b1.1×2 ¹³ = -12288.0 |
| 0x1c = 0.001110.0 = +0b1.0×2 ⁻¹⁸ ≈ 3.8146973E-06 | 0x5c = 0.101110.0 = +0b1.0×2 ¹⁴ = 16384.0 | 0x9c = 1.001110.0 = -0b1.0×2 ⁻¹⁸ ≈ -3.8146973E-06 | 0xdc = 1.101110.0 = -0b1.0×2 ¹⁴ = -16384.0 |
| 0x1d = 0.001110.1 = +0b1.1×2 ⁻¹⁸ ≈ 5.7220459E-06 | 0x5d = 0.101110.1 = +0b1.1×2 ¹⁴ = 24576.0 | 0x9d = 1.001110.1 = -0b1.1×2 ⁻¹⁸ ≈ -5.7220459E-06 | 0xdd = 1.101110.1 = -0b1.1×2 ¹⁴ = -24576.0 |
| 0x1e = 0.001111.0 = +0b1.0×2 ⁻¹⁷ ≈ 7.6293945E-06 | 0x5e = 0.101111.0 = +0b1.0×2 ¹⁵ = 32768.0 | 0x9e = 1.001111.0 = -0b1.0×2 ⁻¹⁷ ≈ -7.6293945E-06 | 0xde = 1.101111.0 = -0b1.0×2 ¹⁵ = -32768.0 |
| 0x1f = 0.001111.1 = +0b1.1×2 ⁻¹⁷ ≈ 1.1444092E-05 | 0x5f = 0.101111.1 = +0b1.1×2 ¹⁵ = 49152.0 | 0x9f = 1.001111.1 = -0b1.1×2 ⁻¹⁷ ≈ -1.1444092E-05 | 0xdf = 1.101111.1 = -0b1.1×2 ¹⁵ = -49152.0 |
| 0x20 = 0.010000.0 = +0b1.0×2 ⁻¹⁶ ≈ 1.5258789E-05 | 0x60 = 0.110000.0 = +0b1.0×2 ¹⁶ = 65536.0 | 0xa0 = 1.010000.0 = -0b1.0×2 ⁻¹⁶ ≈ -1.5258789E-05 | 0xea0 = 1.110000.0 = -0b1.0×2 ¹⁶ = -65536.0 |
| 0x21 = 0.010000.1 = +0b1.1×2 ⁻¹⁶ ≈ 2.2888184E-05 | 0x61 = 0.110000.1 = +0b1.1×2 ¹⁶ = 98304.0 | 0xa1 = 1.010000.1 = -0b1.1×2 ⁻¹⁶ ≈ -2.2888184E-05 | 0xea1 = 1.110000.1 = -0b1.1×2 ¹⁶ = -98304.0 |
| 0x22 = 0.010001.0 = +0b1.0×2 ⁻¹⁵ ≈ 3.0517578E-05 | 0x62 = 0.110001.0 = +0b1.0×2 ¹⁷ = 131072.0 | 0xa2 = 1.010001.0 = -0b1.0×2 ⁻¹⁵ ≈ -3.0517578E-05 | 0xea2 = 1.110001.0 = -0b1.0×2 ¹⁷ = -131072.0 |
| 0x23 = 0.010001.1 = +0b1.1×2 ⁻¹⁵ ≈ 4.5776367E-05 | 0x63 = 0.110001.1 = +0b1.1×2 ¹⁷ = 196608.0 | 0xa3 = 1.010001.1 = -0b1.1×2 ⁻¹⁵ ≈ -4.5776367E-05 | 0xea3 = 1.110001.1 = -0b1.1×2 ¹⁷ = -196608.0 |
| 0x24 = 0.010010.0 = +0b1.0×2 ⁻¹⁴ ≈ 6.1035156E-05 | 0x64 = 0.110010.0 = +0b1.0×2 ¹⁸ = 262144.0 | 0xa4 = 1.010010.0 = -0b1.0×2 ⁻¹⁴ ≈ -6.1035156E-05 | 0xea4 = 1.110010.0 = -0b1.0×2 ¹⁸ = -262144.0 |
| 0x25 = 0.010010.1 = +0b1.1×2 ⁻¹⁴ ≈ 9.1552734E-05 | 0x65 = 0.110010.1 = +0b1.1×2 ¹⁸ = 393216.0 | 0xa5 = 1.010010.1 = -0b1.1×2 ⁻¹⁴ ≈ -9.1552734E-05 | 0xea5 = 1.110010.1 = -0b1.1×2 ¹⁸ = -393216.0 |
| 0x26 = 0.010011.0 = +0b1.0×2 ⁻¹³ ≈ 0.00012207031 | 0x66 = 0.110011.0 = +0b1.0×2 ¹⁹ = 524288.0 | 0xa6 = 1.010011.0 = -0b1.0×2 ⁻¹³ ≈ -0.00012207031 | 0xea6 = 1.110011.0 = -0b1.0×2 ¹⁹ = -524288.0 |
| 0x27 = 0.010011.1 = +0b1.1×2 ⁻¹³ ≈ 0.00018310547 | 0x67 = 0.110011.1 = +0b1.1×2 ¹⁹ = 786432.0 | 0xa7 = 1.010011.1 = -0b1.1×2 ⁻¹³ ≈ -0.00018310547 | 0xea7 = 1.110011.1 = -0b1.1×2 ¹⁹ = -786432.0 |
| 0x28 = 0.010100.0 = +0b1.0×2 ⁻¹² ≈ 0.000244140625 | 0x68 = 0.110100.0 = +0b1.0×2 ²⁰ = 1048576.0 | 0xa8 = 1.010100.0 = -0b1.0×2 ⁻¹² ≈ -0.000244140625 | 0xea8 = 1.110100.0 = -0b1.0×2 ²⁰ = -1048576.0 |
| 0x29 = 0.010100.1 = +0b1.1×2 ⁻¹² ≈ 0.00036621094 | 0x69 = 0.110100.1 = +0b1.1×2 ²⁰ = 1572864.0 | 0xa9 = 1.010100.1 = -0b1.1×2 ⁻¹² ≈ -0.00036621094 | 0xea9 = 1.110100.1 = -0b1.1×2 ²⁰ = -1572864.0 |
| 0x2a = 0.010101.0 = +0b1.0×2 ⁻¹¹ ≈ 0.00048828125 | 0x6a = 0.110101.0 = +0b1.0×2 ²¹ = 2097152.0 | 0xaa = 1.010101.0 = -0b1.0×2 ⁻¹¹ ≈ -0.00048828125 | 0xea = 1.110101.0 = -0b1.0×2 ²¹ = -2097152.0 |
| 0x2b = 0.010101.1 = +0b1.1×2 ⁻¹¹ ≈ 0.000732421875 | 0x6b = 0.110101.1 = +0b1.1×2 ²¹ = 3145728.0 | 0xab = 1.010101.1 = -0b1.1×2 ⁻¹¹ ≈ -0.000732421875 | 0xeb = 1.110101.1 = -0b1.1×2 ²¹ = -3145728.0 |
| 0x2c = 0.010110.0 = +0b1.0×2 ⁻¹⁰ ≈ 0.0009765625 | 0x6c = 0.110110.0 = +0b1.0×2 ²² = 4194304.0 | 0xac = 1.010110.0 = -0b1.0×2 ⁻¹⁰ ≈ -0.0009765625 | 0xec = 1.110110.0 = -0b1.0×2 ²² = -4194304.0 |
| 0x2d = 0.010110.1 = +0b1.1×2 ⁻¹⁰ ≈ 0.00146484375 | 0x6d = 0.110110.1 = +0b1.1×2 ²² = 6291456.0 | 0xad = 1.010110.1 = -0b1.1×2 ⁻¹⁰ ≈ -0.00146484375 | 0xed = 1.110110.1 = -0b1.1×2 ²² = -6291456.0 |
| 0x2e = 0.010111.0 = +0b1.0×2 ⁻⁹ ≈ 0.001953125 | 0x6e = 0.110111.0 = +0b1.0×2 ²³ = 8388608.0 | 0xae = 1.010111.0 = -0b1.0×2 ⁻⁹ ≈ -0.001953125 | 0xee = 1.110111.0 = -0b1.0×2 ²³ = -8388608.0 |
| 0x2f = 0.010111.1 = +0b1.1×2 ⁻⁹ ≈ 0.0029296875 | 0x6f = 0.110111.1 = +0b1.1×2 ²³ = 12582912.0 | 0xaf = 1.010111.1 = -0b1.1×2 ⁻⁹ ≈ -0.0029296875 | 0xef = 1.110111.1 = -0b1.1×2 ²³ = -12582912.0 |
| 0x30 = 0.011000.0 = +0b1.0×2 ⁻⁸ ≈ 0.00390625 | 0x70 = 0.111000.0 = +0b1.0×2 ²⁴ = 16777216.0 | 0xb0 = 1.011000.0 = -0b1.0×2 ⁻⁸ ≈ -0.00390625 | 0xf0 = 1.111000.0 = -0b1.0×2 ²⁴ = -16777216.0 |
| 0x31 = 0.011000.1 = +0b1.1×2 ⁻⁸ ≈ 0.005859375 | 0x71 = 0.111000.1 = +0b1.1×2 ²⁴ = 25165824.0 | 0xb1 = 1.011000.1 = -0b1.1×2 ⁻⁸ ≈ -0.005859375 | 0xf1 = 1.111000.1 = -0b1.1×2 ²⁴ = -25165824.0 |
| 0x32 = 0.011001.0 = +0b1.0×2 ⁻⁷ ≈ 0.0078125 | 0x72 = 0.111001.0 = +0b1.0×2 ²⁵ = 33554432.0 | 0xb2 = 1.011001.0 = -0b1.0×2 ⁻⁷ ≈ -0.0078125 | 0xf2 = 1.111001.0 = -0b1.0×2 ²⁵ = -33554432.0 |
| 0x33 = 0.011001.1 = +0b1.1×2 ⁻⁷ ≈ 0.01171875 | 0x73 = 0.111001.1 = +0b1.1×2 ²⁵ = 50331648.0 | 0xb3 = 1.011001.1 = -0b1.1×2 ⁻⁷ ≈ -0.01171875 | 0xf3 = 1.111001.1 = -0b1.1×2 ²⁵ = -50331648.0 |
| 0x34 = 0.011010.0 = +0b1.0×2 ⁻⁶ ≈ 0.015625 | 0x74 = 0.111010.0 = +0b1.0×2 ²⁶ = 67108864.0 | 0xb4 = 1.011010.0 = -0b1.0×2 ⁻⁶ ≈ -0.015625 | 0xf4 = 1.111010.0 = -0b1.0×2 ²⁶ = -67108864.0 |
| 0x35 = 0.011010.1 = +0b1.1×2 ⁻⁶ ≈ 0.0234375 | 0x75 = 0.111010.1 = +0b1.1×2 ²⁶ = 102466329.6 | 0xb5 = 1.011010.1 = -0b1.1×2 ⁻⁶ ≈ -0.0234375 | 0xf5 = 1.111010.1 = -0b1.1×2 ²⁶ = -102466329.6 |
| 0x36 = 0.011011.0 = +0b1.0×2 ⁻⁵ ≈ 0.03125 | 0x76 = 0.111011.0 = +0b1.0×2 ²⁷ = 134217728.0 | 0xb6 = 1.011011.0 = -0b1.0×2 ⁻⁵ ≈ -0.03125 | 0xf6 = 1.111011.0 = -0b1.0×2 ²⁷ = -134217728.0 |
| 0x37 = 0.011011.1 = +0b1.1×2 ⁻⁵ ≈ 0.046875 | 0x77 = 0.111011.1 = +0b1.1×2 ²⁷ = 201326592.0 | 0xb7 = 1.011011.1 = -0b1.1×2 ⁻⁵ ≈ -0.046875 | 0xf7 = 1.111011.1 = -0b1.1×2 ²⁷ = -201326592.0 |
| 0x38 = 0.011100.0 = +0b1.0×2 ⁻⁴ ≈ 0.0625 | 0x78 = 0.111100.0 = +0b1.0×2 ²⁸ = 268435456.0 | 0xb8 = 1.011100.0 = -0b1.0×2 ⁻⁴ ≈ -0.0625 | 0xf8 = 1.111100.0 = -0b1.0×2 ²⁸ = -268435456.0 |
| 0x39 = 0.011100.1 = +0b1.1×2 ⁻⁴ ≈ 0.09375 | 0x79 = 0.111100.1 = +0b1.1×2 ²⁸ = 402653184.0 | 0xb9 = 1.011100.1 = -0b1.1×2 ⁻⁴ ≈ -0.09375 | 0xf9 = 1.111100.1 = -0b1.1×2 ²⁸ = -402653184.0 |
| 0x3a = 0.011101.0 = +0b1.0×2 ⁻³ ≈ 0.125 | 0x7a = 0.111101.0 = +0b1.0×2 ²⁹ = 536870912.0 | 0xba = 1.011101.0 = -0b1.0×2 ⁻³ ≈ -0.125 | 0xfa = 1.111101.0 = -0b1.0×2 ²⁹ = -536870912.0 |
| 0x3b = 0.011101.1 = +0b1.1×2 ⁻³ ≈ 0.1875 | 0x7b = 0.111101.1 = +0b1.1×2 ²⁹ = 805306368.0 | 0xbb = 1.011101.1 = -0b1.1×2 ⁻³ ≈ -0.1875 | 0xfb = 1.111101.1 = -0b1.1×2 ²⁹ = -805306368.0 |
| 0x3c = 0.011110.0 = +0b1.0×2 ⁻² ≈ 0.25 | 0x7c = 0.111110.0 = +0b1.0×2 ³⁰ = 1073741824.0 | 0xbc = 1.011110.0 = -0b1.0×2 ⁻² ≈ -0.25 | 0xfc = 1.111110.0 = -0b1.0×2 ³⁰ = -1073741824.0 |
| 0x3d = 0.011110.1 = +0b1.1×2 ⁻² ≈ 0.375 | 0x7d = 0.111110.1 = +0b1.1×2 ³⁰ = 1610612736.0 | 0xbd = 1.011110.1 = -0b1.1×2 ⁻² ≈ -0.375 | 0xfd = 1.111110.1 = -0b1.1×2 ³⁰ = -1610612736.0 |
| 0x3e = 0.011111.0 = +0b1.0×2 ⁻¹ ≈ 0 | | | |

C.2 Value Table: P3, $p = 3$, $\text{emax} = 15$

| | | | |
|---|---|--|--|
| 0x00 = 0.00000.00 = 0.0 | 0x40 = 0.10000.00 = +0b1.00×2 ⁰ = 1.0 | 0x80 = 1.00000.00 = NaN | 0xc0 = 1.10000.00 = -0b1.00×2 ⁰ = -1.0 |
| 0x01 = 0.00000.01 = +0b0.01×2 ⁻¹⁵ ≈ 7.6293945E-06 | 0x41 = 0.10000.01 = +0b1.01×2 ⁰ = 1.25 | 0x81 = 1.00000.01 = -0b0.01×2 ⁻¹⁵ ≈ -7.6293945E-06 | 0xc1 = 1.10000.01 = -0b1.01×2 ⁰ = -1.25 |
| 0x02 = 0.00000.10 = +0b0.10×2 ⁻¹⁵ ≈ 1.5258789E-05 | 0x42 = 0.10000.10 = +0b1.10×2 ⁰ = 1.5 | 0x82 = 1.00000.10 = -0b0.10×2 ⁻¹⁵ ≈ -1.5258789E-05 | 0xc2 = 1.10000.10 = -0b1.10×2 ⁰ = -1.5 |
| 0x03 = 0.00000.11 = +0b0.11×2 ⁻¹⁵ ≈ 2.2888184E-05 | 0x43 = 0.10000.11 = +0b1.11×2 ⁰ = 1.75 | 0x83 = 1.00000.11 = -0b0.11×2 ⁻¹⁵ ≈ -2.2888184E-05 | 0xc3 = 1.10000.11 = -0b1.11×2 ⁰ = -1.75 |
| 0x04 = 0.00001.00 = +0b1.00×2 ⁻¹⁵ ≈ 3.0517578E-05 | 0x44 = 0.10001.00 = +0b1.00×2 ¹ = 2.0 | 0x84 = 1.00001.00 = -0b1.00×2 ⁻¹⁵ ≈ -3.0517578E-05 | 0xc4 = 1.10001.00 = -0b1.00×2 ¹ = -2.0 |
| 0x05 = 0.00001.01 = +0b1.01×2 ⁻¹⁵ ≈ 3.8146973E-05 | 0x45 = 0.10001.01 = +0b1.01×2 ¹ = 2.5 | 0x85 = 1.00001.01 = -0b1.01×2 ⁻¹⁵ ≈ -3.8146973E-05 | 0xc5 = 1.10001.01 = -0b1.01×2 ¹ = -2.5 |
| 0x06 = 0.00001.10 = +0b1.10×2 ⁻¹⁵ ≈ 4.5776367E-05 | 0x46 = 0.10001.10 = +0b1.10×2 ¹ = 3.0 | 0x86 = 1.00001.10 = -0b1.10×2 ⁻¹⁵ ≈ -4.5776367E-05 | 0xc6 = 1.10001.10 = -0b1.10×2 ¹ = -3.0 |
| 0x07 = 0.00001.11 = +0b1.11×2 ⁻¹⁵ ≈ 5.3405762E-05 | 0x47 = 0.10001.11 = +0b1.11×2 ¹ = 3.5 | 0x87 = 1.00001.11 = -0b1.11×2 ⁻¹⁵ ≈ -5.3405762E-05 | 0xc7 = 1.10001.11 = -0b1.11×2 ¹ = -3.5 |
| 0x08 = 0.00010.00 = +0b1.00×2 ⁻¹⁴ ≈ 6.1035156E-05 | 0x48 = 0.10010.00 = +0b1.00×2 ² = 4.0 | 0x88 = 1.00010.00 = -0b1.00×2 ⁻¹⁴ ≈ -6.1035156E-05 | 0xc8 = 1.10010.00 = -0b1.00×2 ² = -4.0 |
| 0x09 = 0.00010.01 = +0b1.01×2 ⁻¹⁴ ≈ 7.6293945E-05 | 0x49 = 0.10010.01 = +0b1.01×2 ² = 5.0 | 0x89 = 1.00010.01 = -0b1.01×2 ⁻¹⁴ ≈ -7.6293945E-05 | 0xc9 = 1.10010.01 = -0b1.01×2 ² = -5.0 |
| 0x0a = 0.00010.10 = +0b1.10×2 ⁻¹⁴ ≈ 9.1552734E-05 | 0x4a = 0.10010.10 = +0b1.10×2 ² = 6.0 | 0x8a = 1.00010.10 = -0b1.10×2 ⁻¹⁴ ≈ -9.1552734E-05 | 0xca = 1.10010.10 = -0b1.10×2 ² = -6.0 |
| 0x0b = 0.00010.11 = +0b1.11×2 ⁻¹⁴ ≈ 0.00010681152 | 0x4b = 0.10010.11 = +0b1.11×2 ² = 7.0 | 0x8b = 1.00010.11 = -0b1.11×2 ⁻¹⁴ ≈ -0.00010681152 | 0xcb = 1.10010.11 = -0b1.11×2 ² = -7.0 |
| 0x0c = 0.00011.00 = +0b1.00×2 ⁻¹³ ≈ 0.00012207031 | 0x4c = 0.10011.00 = +0b1.00×2 ³ = 8.0 | 0x8c = 1.00011.00 = -0b1.00×2 ⁻¹³ ≈ -0.00012207031 | 0xcc = 1.10011.00 = -0b1.00×2 ³ = -8.0 |
| 0x0d = 0.00011.01 = +0b1.01×2 ⁻¹³ ≈ 0.00015258789 | 0x4d = 0.10011.01 = +0b1.01×2 ³ = 10.0 | 0x8d = 1.00011.01 = -0b1.01×2 ⁻¹³ ≈ -0.00015258789 | 0xcd = 1.10011.01 = -0b1.01×2 ³ = -10.0 |
| 0x0e = 0.00011.10 = +0b1.10×2 ⁻¹³ ≈ 0.00018310547 | 0x4e = 0.10011.10 = +0b1.10×2 ³ = 12.0 | 0x8e = 1.00011.10 = -0b1.10×2 ⁻¹³ ≈ -0.00018310547 | 0xce = 1.10011.10 = -0b1.10×2 ³ = -12.0 |
| 0x0f = 0.00011.11 = +0b1.11×2 ⁻¹³ ≈ 0.00021362305 | 0x4f = 0.10011.11 = +0b1.11×2 ³ = 14.0 | 0x8f = 1.00011.11 = -0b1.11×2 ⁻¹³ ≈ -0.00021362305 | 0xcf = 1.10011.11 = -0b1.11×2 ³ = -14.0 |
| 0x10 = 0.00100.00 = +0b1.00×2 ⁻¹² = 0.000244140625 | 0x50 = 0.10100.00 = +0b1.00×2 ⁴ = 16.0 | 0x90 = 1.00100.00 = -0b1.00×2 ⁻¹² ≈ -0.000244140625 | 0xd0 = 1.10100.00 = -0b1.00×2 ⁴ = -16.0 |
| 0x11 = 0.00100.01 = +0b1.01×2 ⁻¹² ≈ 0.00030517578 | 0x51 = 0.10100.01 = +0b1.01×2 ⁴ = 20.0 | 0x91 = 1.00100.01 = -0b1.01×2 ⁻¹² ≈ -0.00030517578 | 0xd1 = 1.10100.01 = -0b1.01×2 ⁴ = -20.0 |
| 0x12 = 0.00100.10 = +0b1.10×2 ⁻¹² ≈ 0.00036621094 | 0x52 = 0.10100.10 = +0b1.10×2 ⁴ = 24.0 | 0x92 = 1.00100.10 = -0b1.10×2 ⁻¹² ≈ -0.00036621094 | 0xd2 = 1.10100.10 = -0b1.10×2 ⁴ = -24.0 |
| 0x13 = 0.00100.11 = +0b1.11×2 ⁻¹² ≈ 0.00042724609 | 0x53 = 0.10100.11 = +0b1.11×2 ⁴ = 28.0 | 0x93 = 1.00100.11 = -0b1.11×2 ⁻¹² ≈ -0.00042724609 | 0xd3 = 1.10100.11 = -0b1.11×2 ⁴ = -28.0 |
| 0x14 = 0.00101.00 = +0b1.00×2 ⁻¹¹ ≈ 0.00048828125 | 0x54 = 0.10101.00 = +0b1.00×2 ⁵ = 32.0 | 0x94 = 1.00101.00 = -0b1.00×2 ⁻¹¹ ≈ -0.00048828125 | 0xd4 = 1.10101.00 = -0b1.00×2 ⁵ = -32.0 |
| 0x15 = 0.00101.01 = +0b1.01×2 ⁻¹¹ ≈ 0.00061035156 | 0x55 = 0.10101.01 = +0b1.01×2 ⁵ = 40.0 | 0x95 = 1.00101.01 = -0b1.01×2 ⁻¹¹ ≈ -0.00061035156 | 0xd5 = 1.10101.01 = -0b1.01×2 ⁵ = -40.0 |
| 0x16 = 0.00101.10 = +0b1.10×2 ⁻¹¹ ≈ 0.000732421875 | 0x56 = 0.10101.10 = +0b1.10×2 ⁵ = 48.0 | 0x96 = 1.00101.10 = -0b1.10×2 ⁻¹¹ ≈ -0.000732421875 | 0xd6 = 1.10101.10 = -0b1.10×2 ⁵ = -48.0 |
| 0x17 = 0.00101.11 = +0b1.11×2 ⁻¹¹ ≈ 0.00085449219 | 0x57 = 0.10101.11 = +0b1.11×2 ⁵ = 56.0 | 0x97 = 1.00101.11 = -0b1.11×2 ⁻¹¹ ≈ -0.00085449219 | 0xd7 = 1.10101.11 = -0b1.11×2 ⁵ = -56.0 |
| 0x18 = 0.00110.00 = +0b1.00×2 ⁻¹⁰ = 0.0009765625 | 0x58 = 0.10110.00 = +0b1.00×2 ⁶ = 64.0 | 0x98 = 1.00110.00 = -0b1.00×2 ⁻¹⁰ = -0.0009765625 | 0xd8 = 1.10110.00 = -0b1.00×2 ⁶ = -64.0 |
| 0x19 = 0.00110.01 = +0b1.01×2 ⁻¹⁰ = 0.001220703125 | 0x59 = 0.10110.01 = +0b1.01×2 ⁶ = 80.0 | 0x99 = 1.00110.01 = -0b1.01×2 ⁻¹⁰ ≈ -0.0012207031 | 0xd9 = 1.10110.01 = -0b1.01×2 ⁶ = -80.0 |
| 0x1a = 0.00110.10 = +0b1.01×2 ⁻¹⁰ = 0.00146484375 | 0x5a = 0.10110.10 = +0b1.01×2 ⁶ = 96.0 | 0x9a = 1.00110.10 = -0b1.01×2 ⁻¹⁰ ≈ -0.00146484375 | 0xda = 1.10110.10 = -0b1.01×2 ⁶ = -96.0 |
| 0x1b = 0.00110.11 = +0b1.11×2 ⁻¹⁰ = 0.001708984375 | 0x5b = 0.10110.11 = +0b1.11×2 ⁶ = 112.0 | 0x9b = 1.00110.11 = -0b1.11×2 ⁻¹⁰ ≈ -0.0017089844 | 0xdb = 1.10110.11 = -0b1.11×2 ⁶ = -112.0 |
| 0x1c = 0.00111.00 = +0b1.00×2 ⁻⁹ = 0.001953125 | 0x5c = 0.10111.00 = +0b1.00×2 ⁷ = 128.0 | 0x9c = 1.00111.00 = -0b1.00×2 ⁻⁹ = -0.001953125 | 0xdc = 1.10111.00 = -0b1.00×2 ⁷ = -128.0 |
| 0x1d = 0.00111.01 = +0b1.01×2 ⁻⁹ = 0.00244140625 | 0x5d = 0.10111.01 = +0b1.01×2 ⁷ = 160.0 | 0x9d = 1.00111.01 = -0b1.01×2 ⁻⁹ = -0.00244140625 | 0xdd = 1.10111.01 = -0b1.01×2 ⁷ = -160.0 |
| 0x1e = 0.00111.10 = +0b1.10×2 ⁻⁹ = 0.0029296875 | 0x5e = 0.10111.10 = +0b1.10×2 ⁷ = 192.0 | 0x9e = 1.00111.10 = -0b1.10×2 ⁻⁹ = -0.0029296875 | 0xde = 1.10111.10 = -0b1.10×2 ⁷ = -192.0 |
| 0x1f = 0.00111.11 = +0b1.11×2 ⁻⁹ = 0.00341796875 | 0x5f = 0.10111.11 = +0b1.11×2 ⁷ = 224.0 | 0x9f = 1.00111.11 = -0b1.11×2 ⁻⁹ = -0.00341796875 | 0xdf = 1.10111.11 = -0b1.11×2 ⁷ = -224.0 |
| 0x20 = 0.01000.00 = +0b1.00×2 ⁻⁸ = 0.00390625 | 0x60 = 0.11000.00 = +0b1.00×2 ⁸ = 256.0 | 0xa0 = 1.01000.00 = -0b1.00×2 ⁻⁸ = -0.00390625 | 0xe0 = 1.11000.00 = -0b1.00×2 ⁸ = -256.0 |
| 0x21 = 0.01000.01 = +0b1.01×2 ⁻⁸ = 0.0048828125 | 0x61 = 0.11000.01 = +0b1.01×2 ⁸ = 320.0 | 0xa1 = 1.01000.01 = -0b1.01×2 ⁻⁸ = -0.0048828125 | 0xe1 = 1.11000.01 = -0b1.01×2 ⁸ = -320.0 |
| 0x22 = 0.01000.10 = +0b1.10×2 ⁻⁸ = 0.005859375 | 0x62 = 0.11000.10 = +0b1.10×2 ⁸ = 384.0 | 0xa2 = 1.01000.10 = -0b1.10×2 ⁻⁸ = -0.005859375 | 0xe2 = 1.11000.10 = -0b1.10×2 ⁸ = -384.0 |
| 0x23 = 0.01000.11 = +0b1.11×2 ⁻⁸ = 0.0068359375 | 0x63 = 0.11000.11 = +0b1.11×2 ⁸ = 448.0 | 0xa3 = 1.01000.11 = -0b1.11×2 ⁻⁸ = -0.0068359375 | 0xe3 = 1.11000.11 = -0b1.11×2 ⁸ = -448.0 |
| 0x24 = 0.01001.00 = +0b1.00×2 ⁻⁷ = 0.0078125 | 0x64 = 0.11001.00 = +0b1.00×2 ⁹ = 512.0 | 0xa4 = 1.01001.00 = -0b1.00×2 ⁻⁷ = -0.0078125 | 0xe4 = 1.11001.00 = -0b1.00×2 ⁹ = -512.0 |
| 0x25 = 0.01001.01 = +0b1.01×2 ⁻⁷ = 0.009765625 | 0x65 = 0.11001.01 = +0b1.01×2 ⁹ = 640.0 | 0xa5 = 1.01001.01 = -0b1.01×2 ⁻⁷ = -0.009765625 | 0xe5 = 1.11001.01 = -0b1.01×2 ⁹ = -640.0 |
| 0x26 = 0.01001.10 = +0b1.10×2 ⁻⁷ = 0.01171875 | 0x66 = 0.11001.10 = +0b1.10×2 ⁹ = 768.0 | 0xa6 = 1.01001.10 = -0b1.10×2 ⁻⁷ = -0.01171875 | 0xe6 = 1.11001.10 = -0b1.10×2 ⁹ = -768.0 |
| 0x27 = 0.01001.11 = +0b1.11×2 ⁻⁷ = 0.013671875 | 0x67 = 0.11001.11 = +0b1.11×2 ⁹ = 896.0 | 0xa7 = 1.01001.11 = -0b1.11×2 ⁻⁷ = -0.013671875 | 0xe7 = 1.11001.11 = -0b1.11×2 ⁹ = -896.0 |
| 0x28 = 0.01010.00 = +0b1.00×2 ⁻⁶ = 0.015625 | 0x68 = 0.11010.00 = +0b1.00×2 ¹⁰ = 1024.0 | 0xa8 = 1.01010.00 = -0b1.00×2 ⁻⁶ = -0.015625 | 0xe8 = 1.11010.00 = -0b1.00×2 ¹⁰ = -1024.0 |
| 0x29 = 0.01010.01 = +0b1.01×2 ⁻⁶ = 0.01953125 | 0x69 = 0.11010.01 = +0b1.01×2 ¹⁰ = 1280.0 | 0xa9 = 1.01010.01 = -0b1.01×2 ⁻⁶ = -0.01953125 | 0xe9 = 1.11010.01 = -0b1.01×2 ¹⁰ = -1280.0 |
| 0x2a = 0.01010.10 = +0b1.10×2 ⁻⁶ = 0.0234375 | 0x6a = 0.11010.10 = +0b1.10×2 ¹⁰ = 1536.0 | 0xaa = 1.01010.10 = -0b1.10×2 ⁻⁶ = -0.0234375 | 0xea = 1.11010.10 = -0b1.10×2 ¹⁰ = -1536.0 |
| 0x2b = 0.01010.11 = +0b1.11×2 ⁻⁶ = 0.02734375 | 0x6b = 0.11010.11 = +0b1.11×2 ¹⁰ = 1792.0 | 0xab = 1.01010.11 = -0b1.11×2 ⁻⁶ = -0.02734375 | 0xeb = 1.11010.11 = -0b1.11×2 ¹⁰ = -1792.0 |
| 0x2c = 0.01011.00 = +0b1.00×2 ⁻⁵ = 0.03125 | 0x6c = 0.11011.00 = +0b1.00×2 ¹¹ = 2048.0 | 0xac = 1.01011.00 = -0b1.00×2 ⁻⁵ = -0.03125 | 0xec = 1.11011.00 = -0b1.00×2 ¹¹ = -2048.0 |
| 0x2d = 0.01011.01 = +0b1.01×2 ⁻⁵ = 0.0390625 | 0x6d = 0.11011.01 = +0b1.01×2 ¹¹ = 2560.0 | 0xad = 1.01011.01 = -0b1.01×2 ⁻⁵ = -0.0390625 | 0xed = 1.11011.01 = -0b1.01×2 ¹¹ = -2560.0 |
| 0x2e = 0.01011.10 = +0b1.10×2 ⁻⁵ = 0.046875 | 0x6e = 0.11011.10 = +0b1.10×2 ¹¹ = 3072.0 | 0xae = 1.01011.10 = -0b1.10×2 ⁻⁵ = -0.046875 | 0xee = 1.11011.10 = -0b1.10×2 ¹¹ = -3072.0 |
| 0x2f = 0.01011.11 = +0b1.11×2 ⁻⁵ = 0.0546875 | 0x6f = 0.11011.11 = +0b1.11×2 ¹¹ = 3584.0 | 0xaf = 1.01011.11 = -0b1.11×2 ⁻⁵ = -0.0546875 | 0xef = 1.11011.11 = -0b1.11×2 ¹¹ = -3584.0 |
| 0x30 = 0.01100.00 = +0b1.00×2 ⁻⁴ = 0.0625 | 0x70 = 0.11100.00 = +0b1.00×2 ¹² = 4096.0 | 0xb0 = 1.01100.00 = -0b1.00×2 ⁻⁴ = -0.0625 | 0xf0 = 1.11100.00 = -0b1.00×2 ¹² = -4096.0 |
| 0x31 = 0.01100.01 = +0b1.01×2 ⁻⁴ = 0.078125 | 0x71 = 0.11100.01 = +0b1.01×2 ¹² = 5120.0 | 0xb1 = 1.01100.01 = -0b1.01×2 ⁻⁴ = -0.078125 | 0xf1 = 1.11100.01 = -0b1.01×2 ¹² = -5120.0 |
| 0x32 = 0.01100.10 = +0b1.10×2 ⁻⁴ = 0.09375 | 0x72 = 0.11100.10 = +0b1.10×2 ¹² = 6144.0 | 0xb2 = 1.01100.10 = -0b1.10×2 ⁻⁴ = -0.09375 | 0xf2 = 1.11100.10 = -0b1.10×2 ¹² = -6144.0 |
| 0x33 = 0.01100.11 = +0b1.11×2 ⁻⁴ = 0.109375 | 0x73 = 0.11100.11 = +0b1.11×2 ¹² = 7168.0 | 0xb3 = 1.01100.11 = -0b1.11×2 ⁻⁴ = -0.109375 | 0xf3 = 1.11100.11 = -0b1.11×2 ¹² = -7168.0 |
| 0x34 = 0.01101.00 = +0b1.00×2 ⁻³ = 0.125 | 0x74 = 0.11101.00 = +0b1.00×2 ¹³ = 8192.0 | 0xb4 = 1.01101.00 = -0b1.00×2 ⁻³ = -0.125 | 0xf4 = 1.11101.00 = -0b1.00×2 ¹³ = -8192.0 |
| 0x35 = 0.01101.01 = +0b1.01×2 ⁻³ = 0.15625 | 0x75 = 0.11101.01 = +0b1.01×2 ¹³ = 10240.0 | 0xb5 = 1.01101.01 = -0b1.01×2 ⁻³ = -0.15625 | 0xf5 = 1.11101.01 = -0b1.01×2 ¹³ = -10240.0 |
| 0x36 = 0.01101.10 = +0b1.10×2 ⁻³ = 0.1875 | 0x76 = 0.11101.10 = +0b1.10×2 ¹³ = 12288.0 | 0xb6 = 1.01101.10 = -0b1.10×2 ⁻³ = -0.1875 | 0xf6 = 1.11101.10 = -0b1.10×2 ¹³ = -12288.0 |
| 0x37 = 0.01101.11 = +0b1.11×2 ⁻³ = 0.21875 | 0x77 = 0.11101.11 = +0b1.11×2 ¹³ = 14336.0 | 0xb7 = 1.01101.11 = -0b1.11×2 ⁻³ = -0.21875 | 0xf7 = 1.11101.11 = -0b1.11×2 ¹³ = -14336.0 |
| 0x38 = 0.01110.00 = +0b1.00×2 ⁻² = 0.25 | 0x78 = 0.11110.00 = +0b1.00×2 ¹⁴ = 16384.0 | 0xb8 = 1.01110.00 = -0b1.00×2 ⁻² = -0.25 | 0xf8 = 1.11110.00 = -0b1.11×2 ¹⁴ = -16384.0 |
| 0x39 = 0.01110.01 = +0b1.01×2 ⁻² = 0.3125 | 0x79 = 0.11110.01 = +0b1.01×2 ¹⁴ = 20480.0 | 0xb9 = 1.01110.01 = -0b1.01×2 ⁻² = -0.3125 | 0xf9 = 1.11110.01 = -0b1.01×2 ¹⁴ = -20480.0 |
| 0x3a = 0.01110.10 = +0b1.10×2 ⁻² = 0.375 | 0x7a = 0.11110.10 = +0b1.10×2 ¹⁴ = 24576.0 | 0xba = 1.01110.10 = -0b1.10×2 ⁻² = -0.375 | 0xfa = 1.11110.10 = -0b1.10×2 ¹⁴ = -24576.0 |
| 0x3b = 0.01110.11 = +0b1.11×2 ⁻² = 0.4375 | 0x7b = 0.11110.11 = +0b1.11×2 ¹⁴ = 28672.0 | 0xbb = 1.01110.11 = -0b1.11×2 ⁻² = -0.4375 | 0xfb = 1.11110.11 = -0b1.11×2 ¹⁴ = -28672.0 |
| 0x3c = 0.01111.00 = +0b1.00×2 ⁻¹ = 0.5 | 0x7c = 0.11111.00 = +0b1.00×2 ¹⁵ = 32768.0 | 0xbc = 1.01111.00 = -0b1.00×2 ⁻¹ = -0.5 | 0xfc = 1.11111.00 = -0b1.00×2 ¹⁵ = -32768.0 |
| 0x3d = 0.01111.01 = +0b1.01×2 ⁻¹ = 0.625 | 0x7d = 0.11111.01 = +0b1.01×2 ¹⁵ = 40960.0 | 0xbd = 1.01111.01 = -0b1.01×2 ⁻¹ = -0.625 | 0xfd = 1.11111.01 = -0b1.01×2 ¹⁵ = -40960.0 |
| 0x3e = 0.01111.10 = +0b1.10×2 ⁻¹ = 0.75 | 0x7e = 0.11111.10 = +0b1.10×2 ¹⁵ = 49152.0 | 0xbe = 1.01111.10 = -0b1.10×2 ⁻¹ = -0.75 | 0xfe = 1.11111.10 = -0b1.10×2 ¹⁵ = -49152.0 |
| 0x3f = 0.01111.11 = +0b1.11×2 ⁻¹ = 0.875 | 0x7f = 0.11111.11 = +Inf | 0xbf = 1.01111.11 = -0b1.11×2 ⁻¹ = -0.875 | |

C.3 Value Table: P4, $p = 4$, $\text{emax} = 7$

| | | | |
|---|---|--|--|
| $0x00 = 0.0000.000 = 0.0$ | $0x40 = 0.1000.000 = +0b1.000 \times 2^0 = 1.0$ | $0x80 = 1.0000.000 = \text{NaN}$ | $0xc0 = 1.1000.000 = -0b1.000 \times 2^0 = -1.0$ |
| $0x01 = 0.0000.001 = +0b0.001 \times 2^{-7} = 0.0009765625$ | $0x41 = 0.1000.001 = +0b1.001 \times 2^0 = 1.125$ | $0x81 = 1.0000.001 = -0b0.001 \times 2^{-7} = -0.0009765625$ | $0xc1 = 1.1000.001 = -0b1.001 \times 2^0 = -1.125$ |
| $0x02 = 0.0000.010 = +0b0.010 \times 2^{-7} = 0.001953125$ | $0x42 = 0.1000.010 = +0b1.010 \times 2^0 = 1.25$ | $0x82 = 1.0000.010 = -0b0.010 \times 2^{-7} = -0.001953125$ | $0xc2 = 1.1000.010 = -0b1.010 \times 2^0 = -1.25$ |
| $0x03 = 0.0000.011 = +0b0.011 \times 2^{-7} = 0.0029296875$ | $0x43 = 0.1000.011 = +0b1.011 \times 2^0 = 1.375$ | $0x83 = 1.0000.011 = -0b0.011 \times 2^{-7} = -0.0029296875$ | $0xc3 = 1.1000.011 = -0b1.011 \times 2^0 = -1.375$ |
| $0x04 = 0.0000.100 = +0b0.100 \times 2^{-7} = 0.00390625$ | $0x44 = 0.1000.100 = +0b1.100 \times 2^0 = 1.5$ | $0x84 = 1.0000.100 = -0b0.100 \times 2^{-7} = -0.00390625$ | $0xc4 = 1.1000.100 = -0b1.100 \times 2^0 = -1.5$ |
| $0x05 = 0.0000.101 = +0b0.101 \times 2^{-7} = 0.0048828125$ | $0x45 = 0.1000.101 = +0b1.101 \times 2^0 = 1.625$ | $0x85 = 1.0000.101 = -0b0.101 \times 2^{-7} = -0.0048828125$ | $0xc5 = 1.1000.101 = -0b1.101 \times 2^0 = -1.625$ |
| $0x06 = 0.0000.110 = +0b0.110 \times 2^{-7} = 0.005859375$ | $0x46 = 0.1000.110 = +0b1.110 \times 2^0 = 1.75$ | $0x86 = 1.0000.110 = -0b0.110 \times 2^{-7} = -0.005859375$ | $0xc6 = 1.1000.110 = -0b1.110 \times 2^0 = -1.75$ |
| $0x07 = 0.0000.111 = +0b0.111 \times 2^{-7} = 0.0068359375$ | $0x47 = 0.1000.111 = +0b1.111 \times 2^0 = 1.875$ | $0x87 = 1.0000.111 = -0b0.111 \times 2^{-7} = -0.0068359375$ | $0xc7 = 1.1000.111 = -0b1.111 \times 2^0 = -1.875$ |
| $0x08 = 0.0001.000 = +0b1.000 \times 2^{-6} = 0.0078125$ | $0x48 = 0.1001.000 = +0b1.000 \times 2^1 = 2.0$ | $0x88 = 1.0001.000 = -0b1.000 \times 2^{-6} = -0.0078125$ | $0xc8 = 1.1001.000 = -0b1.000 \times 2^1 = -2.0$ |
| $0x09 = 0.0001.001 = +0b1.001 \times 2^{-6} = 0.0087890625$ | $0x49 = 0.1001.001 = +0b1.001 \times 2^1 = 2.25$ | $0x89 = 1.0001.001 = -0b1.001 \times 2^{-6} = -0.0087890625$ | $0xc9 = 1.1001.001 = -0b1.001 \times 2^1 = -2.25$ |
| $0x0a = 0.0001.010 = +0b1.010 \times 2^{-6} = 0.009765625$ | $0x4a = 0.1001.010 = +0b1.010 \times 2^1 = 2.5$ | $0x8a = 1.0001.010 = -0b1.010 \times 2^{-6} = -0.009765625$ | $0xca = 1.1001.010 = -0b1.010 \times 2^1 = -2.5$ |
| $0x0b = 0.0001.011 = +0b1.011 \times 2^{-6} = 0.0107421875$ | $0x4b = 0.1001.011 = +0b1.011 \times 2^1 = 2.75$ | $0x8b = 1.0001.011 = -0b1.011 \times 2^{-6} = -0.0107421875$ | $0xcb = 1.1001.011 = -0b1.011 \times 2^1 = -2.75$ |
| $0x0c = 0.0001.100 = +0b1.100 \times 2^{-6} = 0.01171875$ | $0x4c = 0.1001.100 = +0b1.100 \times 2^1 = 3.0$ | $0x8c = 1.0001.100 = -0b1.100 \times 2^{-6} = -0.01171875$ | $0xcc = 1.1001.100 = -0b1.100 \times 2^1 = -3.0$ |
| $0x0d = 0.0001.101 = +0b1.101 \times 2^{-6} = 0.0126953125$ | $0x4d = 0.1001.101 = +0b1.101 \times 2^1 = 3.25$ | $0x8d = 1.0001.101 = -0b1.101 \times 2^{-6} = -0.0126953125$ | $0xcd = 1.1001.101 = -0b1.101 \times 2^1 = -3.25$ |
| $0x0e = 0.0001.110 = +0b1.110 \times 2^{-6} = 0.013671875$ | $0x4e = 0.1001.110 = +0b1.110 \times 2^1 = 3.5$ | $0x8e = 1.0001.110 = -0b1.110 \times 2^{-6} = -0.013671875$ | $0xce = 1.1001.110 = -0b1.110 \times 2^1 = -3.5$ |
| $0x0f = 0.0001.111 = +0b1.111 \times 2^{-6} = 0.0146484375$ | $0x4f = 0.1001.111 = +0b1.111 \times 2^1 = 3.75$ | $0x8f = 1.0001.111 = -0b1.111 \times 2^{-6} = -0.0146484375$ | $0xcf = 1.1001.111 = -0b1.111 \times 2^1 = -3.75$ |
| $0x10 = 0.0010.000 = +0b1.000 \times 2^{-5} = 0.015625$ | $0x50 = 0.1010.000 = +0b1.000 \times 2^2 = 4.0$ | $0x90 = 1.0010.000 = -0b1.000 \times 2^{-5} = -0.015625$ | $0xd0 = 1.1010.000 = -0b1.000 \times 2^2 = -4.0$ |
| $0x11 = 0.0010.001 = +0b1.001 \times 2^{-5} = 0.017578125$ | $0x51 = 0.1010.001 = +0b1.001 \times 2^2 = 4.5$ | $0x91 = 1.0010.001 = -0b1.001 \times 2^{-5} = -0.017578125$ | $0xd1 = 1.1010.001 = -0b1.001 \times 2^2 = -4.5$ |
| $0x12 = 0.0010.010 = +0b1.010 \times 2^{-5} = 0.01953125$ | $0x52 = 0.1010.010 = +0b1.010 \times 2^2 = 5.0$ | $0x92 = 1.0010.010 = -0b1.010 \times 2^{-5} = -0.01953125$ | $0xd2 = 1.1010.010 = -0b1.010 \times 2^2 = -5.0$ |
| $0x13 = 0.0010.011 = +0b1.011 \times 2^{-5} = 0.021484375$ | $0x53 = 0.1010.011 = +0b1.011 \times 2^2 = 5.5$ | $0x93 = 1.0010.011 = -0b1.011 \times 2^{-5} = -0.021484375$ | $0xd3 = 1.1010.011 = -0b1.011 \times 2^2 = -5.5$ |
| $0x14 = 0.0010.100 = +0b1.100 \times 2^{-5} = 0.0234375$ | $0x54 = 0.1010.100 = +0b1.100 \times 2^2 = 6.0$ | $0x94 = 1.0010.100 = -0b1.100 \times 2^{-5} = -0.0234375$ | $0xd4 = 1.1010.100 = -0b1.100 \times 2^2 = -6.0$ |
| $0x15 = 0.0010.101 = +0b1.101 \times 2^{-5} = 0.025390625$ | $0x55 = 0.1010.101 = +0b1.101 \times 2^2 = 6.5$ | $0x95 = 1.0010.101 = -0b1.101 \times 2^{-5} = -0.025390625$ | $0xd5 = 1.1010.101 = -0b1.101 \times 2^2 = -6.5$ |
| $0x16 = 0.0010.110 = +0b1.110 \times 2^{-5} = 0.02734375$ | $0x56 = 0.1010.110 = +0b1.110 \times 2^2 = 7.0$ | $0x96 = 1.0010.110 = -0b1.110 \times 2^{-5} = -0.02734375$ | $0xd6 = 1.1010.110 = -0b1.110 \times 2^2 = -7.0$ |
| $0x17 = 0.0010.111 = +0b1.111 \times 2^{-5} = 0.029296875$ | $0x57 = 0.1010.111 = +0b1.111 \times 2^2 = 7.5$ | $0x97 = 1.0010.111 = -0b1.111 \times 2^{-5} = -0.029296875$ | $0xd7 = 1.1010.111 = -0b1.111 \times 2^2 = -7.5$ |
| $0x18 = 0.0011.000 = +0b1.000 \times 2^{-4} = 0.03125$ | $0x58 = 0.1011.000 = +0b1.000 \times 2^3 = 8.0$ | $0x98 = 1.0011.000 = -0b1.000 \times 2^{-4} = -0.03125$ | $0xd8 = 1.1011.000 = -0b1.000 \times 2^3 = -8.0$ |
| $0x19 = 0.0011.001 = +0b1.001 \times 2^{-4} = 0.03515625$ | $0x59 = 0.1011.001 = +0b1.001 \times 2^3 = 9.0$ | $0x99 = 1.0011.001 = -0b1.001 \times 2^{-4} = -0.03515625$ | $0xd9 = 1.1011.001 = -0b1.001 \times 2^3 = -9.0$ |
| $0x1a = 0.0011.010 = +0b1.010 \times 2^{-4} = 0.0390625$ | $0x5a = 0.1011.010 = +0b1.010 \times 2^3 = 10.0$ | $0x9a = 1.0011.010 = -0b1.010 \times 2^{-4} = -0.0390625$ | $0xda = 1.1011.010 = -0b1.010 \times 2^3 = -10.0$ |
| $0x1b = 0.0011.011 = +0b1.011 \times 2^{-4} = 0.04296875$ | $0x5b = 0.1011.011 = +0b1.011 \times 2^3 = 11.0$ | $0x9b = 1.0011.011 = -0b1.011 \times 2^{-4} = -0.04296875$ | $0xdb = 1.1011.011 = -0b1.011 \times 2^3 = -11.0$ |
| $0x1c = 0.0011.100 = +0b1.100 \times 2^{-4} = 0.046875$ | $0x5c = 0.1011.100 = +0b1.100 \times 2^3 = 12.0$ | $0x9c = 1.0011.100 = -0b1.100 \times 2^{-4} = -0.046875$ | $0xdc = 1.1011.100 = -0b1.100 \times 2^3 = -12.0$ |
| $0x1d = 0.0011.101 = +0b1.101 \times 2^{-4} = 0.05078125$ | $0x5d = 0.1011.101 = +0b1.101 \times 2^3 = 13.0$ | $0x9d = 1.0011.101 = -0b1.101 \times 2^{-4} = -0.05078125$ | $0xdd = 1.1011.101 = -0b1.101 \times 2^3 = -13.0$ |
| $0x1e = 0.0011.110 = +0b1.110 \times 2^{-4} = 0.0546875$ | $0x5e = 0.1011.110 = +0b1.110 \times 2^3 = 14.0$ | $0x9e = 1.0011.110 = -0b1.110 \times 2^{-4} = -0.0546875$ | $0xde = 1.1011.110 = -0b1.110 \times 2^3 = -14.0$ |
| $0x1f = 0.0011.111 = +0b1.111 \times 2^{-4} = 0.05859375$ | $0x5f = 0.1011.111 = +0b1.111 \times 2^3 = 15.0$ | $0x9f = 1.0011.111 = -0b1.111 \times 2^{-4} = -0.05859375$ | $0xdf = 1.1011.111 = -0b1.111 \times 2^3 = -15.0$ |
| $0x20 = 0.0100.000 = +0b1.000 \times 2^{-3} = 0.0625$ | $0x60 = 0.1100.000 = +0b1.000 \times 2^4 = 16.0$ | $0xa0 = 1.0100.000 = -0b1.000 \times 2^{-3} = -0.0625$ | $0xe0 = 1.1100.000 = -0b1.000 \times 2^4 = -16.0$ |
| $0x21 = 0.0100.001 = +0b1.001 \times 2^{-3} = 0.0703125$ | $0x61 = 0.1100.001 = +0b1.001 \times 2^4 = 18.0$ | $0xa1 = 1.0100.001 = -0b1.001 \times 2^{-3} = -0.0703125$ | $0xe1 = 1.1100.001 = -0b1.001 \times 2^4 = -18.0$ |
| $0x22 = 0.0100.010 = +0b1.010 \times 2^{-3} = 0.078125$ | $0x62 = 0.1100.010 = +0b1.010 \times 2^4 = 20.0$ | $0xa2 = 1.0100.010 = -0b1.010 \times 2^{-3} = -0.078125$ | $0xe2 = 1.1100.010 = -0b1.010 \times 2^4 = -20.0$ |
| $0x23 = 0.0100.011 = +0b1.011 \times 2^{-3} = 0.0859375$ | $0x63 = 0.1100.011 = +0b1.011 \times 2^4 = 22.0$ | $0xa3 = 1.0100.011 = -0b1.011 \times 2^{-3} = -0.0859375$ | $0xe3 = 1.1100.011 = -0b1.011 \times 2^4 = -22.0$ |
| $0x24 = 0.0100.100 = +0b1.100 \times 2^{-3} = 0.09375$ | $0x64 = 0.1100.100 = +0b1.100 \times 2^4 = 24.0$ | $0xa4 = 1.0100.100 = -0b1.100 \times 2^{-3} = -0.09375$ | $0xe4 = 1.1100.100 = -0b1.100 \times 2^4 = -24.0$ |
| $0x25 = 0.0100.101 = +0b1.101 \times 2^{-3} = 0.1015625$ | $0x65 = 0.1100.101 = +0b1.101 \times 2^4 = 26.0$ | $0xa5 = 1.0100.101 = -0b1.101 \times 2^{-3} = -0.1015625$ | $0xe5 = 1.1100.101 = -0b1.101 \times 2^4 = -26.0$ |
| $0x26 = 0.0100.110 = +0b1.110 \times 2^{-3} = 0.109375$ | $0x66 = 0.1100.110 = +0b1.110 \times 2^4 = 28.0$ | $0xa6 = 1.0100.110 = -0b1.110 \times 2^{-3} = -0.109375$ | $0xe6 = 1.1100.110 = -0b1.110 \times 2^4 = -28.0$ |
| $0x27 = 0.0100.111 = +0b1.111 \times 2^{-3} = 0.1171875$ | $0x67 = 0.1100.111 = +0b1.111 \times 2^4 = 30.0$ | $0xa7 = 1.0100.111 = -0b1.111 \times 2^{-3} = -0.1171875$ | $0xe7 = 1.1100.111 = -0b1.111 \times 2^4 = -30.0$ |
| $0x28 = 0.0101.000 = +0b1.000 \times 2^{-2} = 0.125$ | $0x68 = 0.1101.000 = +0b1.000 \times 2^5 = 32.0$ | $0xa8 = 1.0101.000 = -0b1.000 \times 2^{-2} = -0.125$ | $0xe8 = 1.1101.000 = -0b1.000 \times 2^5 = -32.0$ |
| $0x29 = 0.0101.001 = +0b1.001 \times 2^{-2} = 0.140625$ | $0x69 = 0.1101.001 = +0b1.001 \times 2^5 = 36.0$ | $0xa9 = 1.0101.001 = -0b1.001 \times 2^{-2} = -0.140625$ | $0xe9 = 1.1101.001 = -0b1.001 \times 2^5 = -36.0$ |
| $0x2a = 0.0101.010 = +0b1.010 \times 2^{-2} = 0.15625$ | $0x6a = 0.1101.010 = +0b1.010 \times 2^5 = 40.0$ | $0xaa = 1.0101.010 = -0b1.010 \times 2^{-2} = -0.15625$ | $0xea = 1.1101.010 = -0b1.010 \times 2^5 = -40.0$ |
| $0x2b = 0.0101.011 = +0b1.011 \times 2^{-2} = 0.171875$ | $0x6b = 0.1101.011 = +0b1.011 \times 2^5 = 44.0$ | $0xab = 1.0101.011 = -0b1.011 \times 2^{-2} = -0.171875$ | $0xeb = 1.1101.011 = -0b1.011 \times 2^5 = -44.0$ |
| $0x2c = 0.0101.100 = +0b1.100 \times 2^{-2} = 0.1875$ | $0x6c = 0.1101.100 = +0b1.100 \times 2^5 = 48.0$ | $0xac = 1.0101.100 = -0b1.100 \times 2^{-2} = -0.1875$ | $0xec = 1.1101.100 = -0b1.100 \times 2^5 = -48.0$ |
| $0x2d = 0.0101.101 = +0b1.101 \times 2^{-2} = 0.203125$ | $0x6d = 0.1101.101 = +0b1.101 \times 2^5 = 52.0$ | $0xad = 1.0101.101 = -0b1.101 \times 2^{-2} = -0.203125$ | $0xed = 1.1101.101 = -0b1.101 \times 2^5 = -52.0$ |
| $0x2e = 0.0101.110 = +0b1.110 \times 2^{-2} = 0.21875$ | $0x6e = 0.1101.110 = +0b1.110 \times 2^5 = 56.0$ | $0xae = 1.0101.110 = -0b1.110 \times 2^{-2} = -0.21875$ | $0xee = 1.1101.110 = -0b1.110 \times 2^5 = -56.0$ |
| $0x2f = 0.0101.111 = +0b1.111 \times 2^{-2} = 0.234375$ | $0x6f = 0.1101.111 = +0b1.111 \times 2^5 = 60.0$ | $0xaf = 1.0101.111 = -0b1.111 \times 2^{-2} = -0.234375$ | $0xef = 1.1101.111 = -0b1.111 \times 2^5 = -60.0$ |
| $0x30 = 0.0110.000 = +0b1.000 \times 2^{-1} = 0.25$ | $0x70 = 0.1110.000 = +0b1.000 \times 2^6 = 64.0$ | $0xb0 = 1.0110.000 = -0b1.000 \times 2^{-1} = -0.25$ | $0xf0 = 1.1110.000 = -0b1.000 \times 2^6 = -64.0$ |
| $0x31 = 0.0110.001 = +0b1.001 \times 2^{-1} = 0.28125$ | $0x71 = 0.1110.001 = +0b1.001 \times 2^6 = 72.0$ | $0xb1 = 1.0110.001 = -0b1.001 \times 2^{-1} = -0.28125$ | $0xf1 = 1.1110.001 = -0b1.001 \times 2^6 = -72.0$ |
| $0x32 = 0.0110.010 = +0b1.010 \times 2^{-1} = 0.3125$ | $0x72 = 0.1110.010 = +0b1.010 \times 2^6 = 80.0$ | $0xb2 = 1.0110.010 = -0b1.010 \times 2^{-1} = -0.3125$ | $0xf2 = 1.1110.010 = -0b1.010 \times 2^6 = -80.0$ |
| $0x33 = 0.0110.011 = +0b1.011 \times 2^{-1} = 0.34375$ | $0x73 = 0.1110.011 = +0b1.011 \times 2^6 = 88.0$ | $0xb3 = 1.0110.011 = -0b1.011 \times 2^{-1} = -0.34375$ | $0xf3 = 1.1110.011 = -0b1.011 \times 2^6 = -88.0$ |
| $0x34 = 0.0110.100 = +0b1.100 \times 2^{-1} = 0.375$ | $0x74 = 0.1110.100 = +0b1.100 \times 2^6 = 96.0$ | $0xb4 = 1.0110.100 = -0b1.100 \times 2^{-1} = -0.375$ | $0xf4 = 1.1110.100 = -0b1.100 \times 2^6 = -96.0$ |
| $0x35 = 0.0110.101 = +0b1.101 \times 2^{-1} = 0.40625$ | $0x75 = 0.1110.101 = +0b1.101 \times 2^6 = 104.0$ | $0xb5 = 1.0110.101 = -0b1.101 \times 2^{-1} = -0.40625$ | $0xf5 = 1.1110.101 = -0b1.101 \times 2^6 = -104.0$ |
| $0x36 = 0.0110.110 = +0b1.110 \times 2^{-1} = 0.4375$ | $0x76 = 0.1110.110 = +0b1.110 \times 2^6 = 112.0$ | $0xb6 = 1.0110.110 = -0b1.110 \times 2^{-1} = -0.4375$ | $0xf6 = 1.1110.110 = -0b1.110 \times 2^6 = -112.0$ |
| $0x37 = 0.0110.111 = +0b1.111 \times 2^{-1} = 0.46875$ | $0x77 = 0.1110.111 = +0b1.111 \times 2^6 = 120.0$ | $0xb7 = 1.0110.111 = -0b1.111 \times 2^{-1} = -0.46875$ | $0xf7 = 1.1110.111 = -0b1.111 \times 2^6 = -120.0$ |
| $0x38 = 0.0111.000 = +0b1.000 \times 2^{-1} = 0.5$ | $0x78 = 0.1111.000 = +0b1.000 \times 2^7 = 128.0$ | $0xb8 = 1.0111.000 = -0b1.000 \times 2^{-1} = -0.5$ | $0xf8 = 1.1111.000 = -0b1.000 \times 2^7 = -128.0$ |
| $0x39 = 0.0111.001 = +0b1.001 \times 2^{-1} = 0.5625$ | $0x79 = 0.1111.001 = +0b1.001 \times 2^7 = 144.0$ | $0xb9 = 1.0111.001 = -0b1.001 \times 2^{-1} = -0.5625$ | $0xf9 = 1.1111.001 = -0b1.001 \times 2^7 = -144.0$ |
| $0x3a = 0.0111.010 = +0b1.010 \times 2^{-1} = 0.625$ | $0x7a = 0.1111.010 = +0b1.010 \times 2^7 = 160.0$ | $0xba = 1.0111.010 = -0b1.010 \times 2^{-1} = -0.625$ | $0xfa = 1.1111.010 = -0b1.010 \times 2^7 = -160.0$ |
| $0x3b = 0.0111.011 = +0b1.011 \times 2^{-1} = 0.6875$ | $0x7b = 0.1111.011 = +0b1.011 \times 2^7 = 176.0$ | $0xbb = 1.0111.011 = -0b1.011 \times 2^{-1} = -0.6875$ | $0xfb = 1.1111.011 = -0b1.011 \times 2^7 = -176.0$ |
| $0x3c = 0.0111.100 = +0b1.100 \times 2^{-1} = 0.75$ | $0x7c = 0.1111.100 = +0b1.100 \times 2^7 = 192.0$ | $0xbc = 1.0111.100 = -0b1.100 \times 2^{-1} = -0.75$ | $0xfc = 1.1111.100 = -0b1.100 \times 2^7 = -192.0$ |
| $0x3d = 0.0111.101 = +0b1.101 \times 2^{-1} = 0.8125$ | $0x7d = 0.1111.101 = +0b1.101 \times 2^7 = 208.0$ | $0xbd = 1.0111.101 = -0b1.101 \times 2^{-1} = -0.8125$ | $0xfd = 1.1111.101 = -0b1.101 \times 2^7 = -208.0$ |
| $0x3e = 0.0111.110 = +0b1.110 \times 2^{-1} = 0.875$ | $0x7e = 0.1111.110 = +0b1.110 \times 2^7 = 224.0$ | $0xbe = 1.0111.110 = -0b$ | |

C.4 Value Table: P5, $p = 5$, $\text{emax} = 3$

| | | | |
|---|---|--|--|
| 0x00 = 0.000.0000 = 0.0 | 0x40 = 0.100.0000 = +0b1.0000×2 ⁰ = 1.0 | 0x80 = 1.000.0000 = NaN | 0xc0 = 1.100.0000 = -0b1.0000×2 ⁰ = -1.0 |
| 0x01 = 0.000.0001 = +0b0.0001×2 ⁻³ = 0.0078125 | 0x41 = 0.100.0001 = +0b1.0001×2 ⁰ = 1.0625 | 0x81 = 1.000.0001 = -0b0.0001×2 ⁻³ = -0.0078125 | 0xc1 = 1.100.0001 = -0b1.0001×2 ⁰ = -1.0625 |
| 0x02 = 0.000.0010 = +0b0.0010×2 ⁻³ = 0.015625 | 0x42 = 0.100.0010 = +0b1.0010×2 ⁰ = 1.125 | 0x82 = 1.000.0010 = -0b0.0010×2 ⁻³ = -0.015625 | 0xc2 = 1.100.0010 = -0b1.0010×2 ⁰ = -1.125 |
| 0x03 = 0.000.0011 = +0b0.0011×2 ⁻³ = 0.0234375 | 0x43 = 0.100.0011 = +0b1.0011×2 ⁰ = 1.1875 | 0x83 = 1.000.0011 = -0b0.0011×2 ⁻³ = -0.0234375 | 0xc3 = 1.100.0011 = -0b1.0011×2 ⁰ = -1.1875 |
| 0x04 = 0.000.0100 = +0b0.0100×2 ⁻³ = 0.03125 | 0x44 = 0.100.0100 = +0b1.0100×2 ⁰ = 1.25 | 0x84 = 1.000.0100 = -0b0.0100×2 ⁻³ = -0.03125 | 0xc4 = 1.100.0100 = -0b1.0100×2 ⁰ = -1.25 |
| 0x05 = 0.000.0101 = +0b0.0101×2 ⁻³ = 0.0390625 | 0x45 = 0.100.0101 = +0b1.0101×2 ⁰ = 1.3125 | 0x85 = 1.000.0101 = -0b0.0101×2 ⁻³ = -0.0390625 | 0xc5 = 1.100.0101 = -0b1.0101×2 ⁰ = -1.3125 |
| 0x06 = 0.000.0110 = +0b0.0110×2 ⁻³ = 0.046875 | 0x46 = 0.100.0110 = +0b1.0110×2 ⁰ = 1.375 | 0x86 = 1.000.0110 = -0b0.0110×2 ⁻³ = -0.046875 | 0xc6 = 1.100.0110 = -0b1.0110×2 ⁰ = -1.375 |
| 0x07 = 0.000.0111 = +0b0.0111×2 ⁻³ = 0.0546875 | 0x47 = 0.100.0111 = +0b1.0111×2 ⁰ = 1.4375 | 0x87 = 1.000.0111 = -0b0.0111×2 ⁻³ = -0.0546875 | 0xc7 = 1.100.0111 = -0b1.0111×2 ⁰ = -1.4375 |
| 0x08 = 0.000.1000 = +0b0.1000×2 ⁻³ = 0.0625 | 0x48 = 0.100.1000 = +0b1.1000×2 ⁰ = 1.5 | 0x88 = 1.000.1000 = -0b0.1000×2 ⁻³ = -0.0625 | 0xc8 = 1.100.1000 = -0b1.1000×2 ⁰ = -1.5 |
| 0x09 = 0.000.1001 = +0b0.1001×2 ⁻³ = 0.0703125 | 0x49 = 0.100.1001 = +0b1.1001×2 ⁰ = 1.5625 | 0x89 = 1.000.1001 = -0b0.1001×2 ⁻³ = -0.0703125 | 0xc9 = 1.100.1001 = -0b1.1001×2 ⁰ = -1.5625 |
| 0x0a = 0.000.1010 = +0b0.1010×2 ⁻³ = 0.078125 | 0x4a = 0.100.1010 = +0b1.1010×2 ⁰ = 1.625 | 0x8a = 1.000.1010 = -0b0.1010×2 ⁻³ = -0.078125 | 0xca = 1.100.1010 = -0b1.1010×2 ⁰ = -1.625 |
| 0x0b = 0.000.1011 = +0b0.1011×2 ⁻³ = 0.0859375 | 0x4b = 0.100.1011 = +0b1.1011×2 ⁰ = 1.6875 | 0x8b = 1.000.1011 = -0b0.1011×2 ⁻³ = -0.0859375 | 0xcb = 1.100.1011 = -0b1.1011×2 ⁰ = -1.6875 |
| 0x0c = 0.000.1100 = +0b0.1100×2 ⁻³ = 0.09375 | 0x4c = 0.100.1100 = +0b1.1100×2 ⁰ = 1.75 | 0x8c = 1.000.1100 = -0b0.1100×2 ⁻³ = -0.09375 | 0xcc = 1.100.1100 = -0b1.1100×2 ⁰ = -1.75 |
| 0x0d = 0.000.1101 = +0b0.1101×2 ⁻³ = 0.1015625 | 0x4d = 0.100.1101 = +0b1.1101×2 ⁰ = 1.8125 | 0x8d = 1.000.1101 = -0b0.1101×2 ⁻³ = -0.1015625 | 0xcd = 1.100.1101 = -0b1.1101×2 ⁰ = -1.8125 |
| 0x0e = 0.000.1110 = +0b0.1110×2 ⁻³ = 0.109375 | 0x4e = 0.100.1110 = +0b1.1110×2 ⁰ = 1.875 | 0x8e = 1.000.1110 = -0b0.1110×2 ⁻³ = -0.109375 | 0xce = 1.100.1110 = -0b1.1110×2 ⁰ = -1.875 |
| 0x0f = 0.000.1111 = +0b0.1111×2 ⁻³ = 0.1171875 | 0x4f = 0.100.1111 = +0b1.1111×2 ⁰ = 1.9375 | 0x8f = 1.000.1111 = -0b0.1111×2 ⁻³ = -0.1171875 | 0xcf = 1.100.1111 = -0b1.1111×2 ⁰ = -1.9375 |
| 0x10 = 0.001.0000 = +0b1.0000×2 ⁻³ = 0.125 | 0x50 = 0.101.0000 = +0b1.0000×2 ¹ = 2.0 | 0x90 = 1.001.0000 = -0b1.0000×2 ⁻³ = -0.125 | 0xd0 = 1.101.0000 = -0b1.0000×2 ¹ = -2.0 |
| 0x11 = 0.001.0001 = +0b1.0001×2 ⁻³ = 0.1328125 | 0x51 = 0.101.0001 = +0b1.0001×2 ¹ = 2.125 | 0x91 = 1.001.0001 = -0b1.0001×2 ⁻³ = -0.1328125 | 0xd1 = 1.101.0001 = -0b1.0001×2 ¹ = -2.125 |
| 0x12 = 0.001.0010 = +0b1.0010×2 ⁻³ = 0.140625 | 0x52 = 0.101.0010 = +0b1.0010×2 ¹ = 2.25 | 0x92 = 1.001.0010 = -0b1.0010×2 ⁻³ = -0.140625 | 0xd2 = 1.101.0010 = -0b1.0010×2 ¹ = -2.25 |
| 0x13 = 0.001.0011 = +0b1.0011×2 ⁻³ = 0.1484375 | 0x53 = 0.101.0011 = +0b1.0011×2 ¹ = 2.375 | 0x93 = 1.001.0011 = -0b1.0011×2 ⁻³ = -0.1484375 | 0xd3 = 1.101.0011 = -0b1.0011×2 ¹ = -2.375 |
| 0x14 = 0.001.0100 = +0b1.0100×2 ⁻³ = 0.15625 | 0x54 = 0.101.0100 = +0b1.0100×2 ¹ = 2.5 | 0x94 = 1.001.0100 = -0b1.0100×2 ⁻³ = -0.15625 | 0xd4 = 1.101.0100 = -0b1.0100×2 ¹ = -2.5 |
| 0x15 = 0.001.0101 = +0b1.0101×2 ⁻³ = 0.1640625 | 0x55 = 0.101.0101 = +0b1.0101×2 ¹ = 2.625 | 0x95 = 1.001.0101 = -0b1.0101×2 ⁻³ = -0.1640625 | 0xd5 = 1.101.0101 = -0b1.0101×2 ¹ = -2.625 |
| 0x16 = 0.001.0110 = +0b1.0110×2 ⁻³ = 0.171875 | 0x56 = 0.101.0110 = +0b1.0110×2 ¹ = 2.75 | 0x96 = 1.001.0110 = -0b1.0110×2 ⁻³ = -0.171875 | 0xd6 = 1.101.0110 = -0b1.0110×2 ¹ = -2.75 |
| 0x17 = 0.001.0111 = +0b1.0111×2 ⁻³ = 0.1796875 | 0x57 = 0.101.0111 = +0b1.0111×2 ¹ = 2.875 | 0x97 = 1.001.0111 = -0b1.0111×2 ⁻³ = -0.1796875 | 0xd7 = 1.101.0111 = -0b1.0111×2 ¹ = -2.875 |
| 0x18 = 0.001.1000 = +0b1.1000×2 ⁻³ = 0.1875 | 0x58 = 0.101.1000 = +0b1.1000×2 ¹ = 3.0 | 0x98 = 1.001.1000 = -0b1.1000×2 ⁻³ = -0.1875 | 0xd8 = 1.101.1000 = -0b1.1000×2 ¹ = -3.0 |
| 0x19 = 0.001.1001 = +0b1.1001×2 ⁻³ = 0.1953125 | 0x59 = 0.101.1001 = +0b1.1001×2 ¹ = 3.125 | 0x99 = 1.001.1001 = -0b1.1001×2 ⁻³ = -0.1953125 | 0xd9 = 1.101.1001 = -0b1.1001×2 ¹ = -3.125 |
| 0x1a = 0.001.1010 = +0b1.1010×2 ⁻³ = 0.203125 | 0x5a = 0.101.1010 = +0b1.1010×2 ¹ = 3.25 | 0x9a = 1.001.1010 = -0b1.1010×2 ⁻³ = -0.203125 | 0xda = 1.101.1010 = -0b1.1010×2 ¹ = -3.25 |
| 0x1b = 0.001.1011 = +0b1.1011×2 ⁻³ = 0.2109375 | 0x5b = 0.101.1011 = +0b1.1011×2 ¹ = 3.375 | 0x9b = 1.001.1011 = -0b1.1011×2 ⁻³ = -0.2109375 | 0xdb = 1.101.1011 = -0b1.1011×2 ¹ = -3.375 |
| 0x1c = 0.001.1100 = +0b1.1100×2 ⁻³ = 0.21875 | 0x5c = 0.101.1100 = +0b1.1100×2 ¹ = 3.5 | 0x9c = 1.001.1100 = -0b1.1100×2 ⁻³ = -0.21875 | 0xdc = 1.101.1100 = -0b1.1100×2 ¹ = -3.5 |
| 0x1d = 0.001.1101 = +0b1.1101×2 ⁻³ = 0.2265625 | 0x5d = 0.101.1101 = +0b1.1101×2 ¹ = 3.625 | 0x9d = 1.001.1101 = -0b1.1101×2 ⁻³ = -0.2265625 | 0xdd = 1.101.1101 = -0b1.1101×2 ¹ = -3.625 |
| 0x1e = 0.001.1110 = +0b1.1110×2 ⁻³ = 0.234375 | 0x5e = 0.101.1110 = +0b1.1110×2 ¹ = 3.75 | 0x9e = 1.001.1110 = -0b1.1110×2 ⁻³ = -0.234375 | 0xde = 1.101.1110 = -0b1.1110×2 ¹ = -3.75 |
| 0x1f = 0.001.1111 = +0b1.1111×2 ⁻³ = 0.2421875 | 0x5f = 0.101.1111 = +0b1.1111×2 ¹ = 3.875 | 0x9f = 1.001.1111 = -0b1.1111×2 ⁻³ = -0.2421875 | 0xdf = 1.101.1111 = -0b1.1111×2 ¹ = -3.875 |
| 0x20 = 0.010.0000 = +0b1.0000×2 ⁻² = 0.25 | 0x60 = 0.110.0000 = +0b1.0000×2 ² = 4.0 | 0xa0 = 1.010.0000 = -0b1.0000×2 ⁻² = -0.25 | 0xe0 = 1.110.0000 = -0b1.0000×2 ² = -4.0 |
| 0x21 = 0.010.0001 = +0b1.0001×2 ⁻² = 0.265625 | 0x61 = 0.110.0001 = +0b1.0001×2 ² = 4.25 | 0xa1 = 1.010.0001 = -0b1.0001×2 ⁻² = -0.265625 | 0xe1 = 1.110.0001 = -0b1.0001×2 ² = -4.25 |
| 0x22 = 0.010.0010 = +0b1.0010×2 ⁻² = 0.28125 | 0x62 = 0.110.0010 = +0b1.0010×2 ² = 4.5 | 0xa2 = 1.010.0010 = -0b1.0010×2 ⁻² = -0.28125 | 0xe2 = 1.110.0010 = -0b1.0010×2 ² = -4.5 |
| 0x23 = 0.010.0011 = +0b1.0011×2 ⁻² = 0.296875 | 0x63 = 0.110.0011 = +0b1.0011×2 ² = 4.75 | 0xa3 = 1.010.0011 = -0b1.0011×2 ⁻² = -0.296875 | 0xe3 = 1.110.0011 = -0b1.0011×2 ² = -4.75 |
| 0x24 = 0.010.0100 = +0b1.0100×2 ⁻² = 0.3125 | 0x64 = 0.110.0100 = +0b1.0100×2 ² = 5.0 | 0xa4 = 1.010.0100 = -0b1.0100×2 ⁻² = -0.3125 | 0xe4 = 1.110.0100 = -0b1.0100×2 ² = -5.0 |
| 0x25 = 0.010.0101 = +0b1.0101×2 ⁻² = 0.328125 | 0x65 = 0.110.0101 = +0b1.0101×2 ² = 5.25 | 0xa5 = 1.010.0101 = -0b1.0101×2 ⁻² = -0.328125 | 0xe5 = 1.110.0101 = -0b1.0101×2 ² = -5.25 |
| 0x26 = 0.010.0110 = +0b1.0110×2 ⁻² = 0.34375 | 0x66 = 0.110.0110 = +0b1.0110×2 ² = 5.5 | 0xa6 = 1.010.0110 = -0b1.0110×2 ⁻² = -0.34375 | 0xe6 = 1.110.0110 = -0b1.0110×2 ² = -5.5 |
| 0x27 = 0.010.0111 = +0b1.0111×2 ⁻² = 0.359375 | 0x67 = 0.110.0111 = +0b1.0111×2 ² = 5.75 | 0xa7 = 1.010.0111 = -0b1.0111×2 ⁻² = -0.359375 | 0xe7 = 1.110.0111 = -0b1.0111×2 ² = -5.75 |
| 0x28 = 0.010.1000 = +0b1.1000×2 ⁻² = 0.375 | 0x68 = 0.110.1000 = +0b1.1000×2 ² = 6.0 | 0xa8 = 1.010.1000 = -0b1.1000×2 ⁻² = -0.375 | 0xe8 = 1.110.1000 = -0b1.1000×2 ² = -6.0 |
| 0x29 = 0.010.1001 = +0b1.1001×2 ⁻² = 0.390625 | 0x69 = 0.110.1001 = +0b1.1001×2 ² = 6.25 | 0xa9 = 1.010.1001 = -0b1.1001×2 ⁻² = -0.390625 | 0xe9 = 1.110.1001 = -0b1.1001×2 ² = -6.25 |
| 0x2a = 0.010.1010 = +0b1.1010×2 ⁻² = 0.40625 | 0x6a = 0.110.1010 = +0b1.1010×2 ² = 6.5 | 0xaa = 1.010.1010 = -0b1.1010×2 ⁻² = -0.40625 | 0xea = 1.110.1010 = -0b1.1010×2 ² = -6.5 |
| 0x2b = 0.010.1011 = +0b1.1011×2 ⁻² = 0.421875 | 0x6b = 0.110.1011 = +0b1.1011×2 ² = 6.75 | 0xab = 1.010.1011 = -0b1.1011×2 ⁻² = -0.421875 | 0xeb = 1.110.1011 = -0b1.1011×2 ² = -6.75 |
| 0x2c = 0.010.1100 = +0b1.1100×2 ⁻² = 0.4375 | 0x6c = 0.110.1100 = +0b1.1100×2 ² = 7.0 | 0xac = 1.010.1100 = -0b1.1100×2 ⁻² = -0.4375 | 0xec = 1.110.1100 = -0b1.1100×2 ² = -7.0 |
| 0x2d = 0.010.1101 = +0b1.1101×2 ⁻² = 0.453125 | 0x6d = 0.110.1101 = +0b1.1101×2 ² = 7.25 | 0xad = 1.010.1101 = -0b1.1101×2 ⁻² = -0.453125 | 0xed = 1.110.1101 = -0b1.1101×2 ² = -7.25 |
| 0x2e = 0.010.1110 = +0b1.1110×2 ⁻² = 0.46875 | 0x6e = 0.110.1110 = +0b1.1110×2 ² = 7.5 | 0xae = 1.010.1110 = -0b1.1110×2 ⁻² = -0.46875 | 0xee = 1.110.1110 = -0b1.1110×2 ² = -7.5 |
| 0x2f = 0.010.1111 = +0b1.1111×2 ⁻² = 0.484375 | 0x6f = 0.110.1111 = +0b1.1111×2 ² = 7.75 | 0xaf = 1.010.1111 = -0b1.1111×2 ⁻² = -0.484375 | 0xef = 1.110.1111 = -0b1.1111×2 ² = -7.75 |
| 0x30 = 0.011.0000 = +0b1.0000×2 ⁻¹ = 0.5 | 0x70 = 0.111.0000 = +0b1.0000×2 ³ = 8.0 | 0xb0 = 1.011.0000 = -0b1.0000×2 ⁻¹ = -0.5 | 0xf0 = 1.111.0000 = -0b1.0000×2 ³ = -8.0 |
| 0x31 = 0.011.0001 = +0b1.0001×2 ⁻¹ = 0.53125 | 0x71 = 0.111.0001 = +0b1.0001×2 ³ = 8.5 | 0xb1 = 1.011.0001 = -0b1.0001×2 ⁻¹ = -0.53125 | 0xf1 = 1.111.0001 = -0b1.0001×2 ³ = -8.5 |
| 0x32 = 0.011.0010 = +0b1.0010×2 ⁻¹ = 0.5625 | 0x72 = 0.111.0010 = +0b1.0010×2 ³ = 9.0 | 0xb2 = 1.011.0010 = -0b1.0010×2 ⁻¹ = -0.5625 | 0xf2 = 1.111.0010 = -0b1.0010×2 ³ = -9.0 |
| 0x33 = 0.011.0011 = +0b1.0011×2 ⁻¹ = 0.59375 | 0x73 = 0.111.0011 = +0b1.0011×2 ³ = 9.5 | 0xb3 = 1.011.0011 = -0b1.0011×2 ⁻¹ = -0.59375 | 0xf3 = 1.111.0011 = -0b1.0011×2 ³ = -9.5 |
| 0x34 = 0.011.0100 = +0b1.0100×2 ⁻¹ = 0.625 | 0x74 = 0.111.0100 = +0b1.0100×2 ³ = 10.0 | 0xb4 = 1.011.0100 = -0b1.0100×2 ⁻¹ = -0.625 | 0xf4 = 1.111.0100 = -0b1.0100×2 ³ = -10.0 |
| 0x35 = 0.011.0101 = +0b1.0101×2 ⁻¹ = 0.65625 | 0x75 = 0.111.0101 = +0b1.0101×2 ³ = 10.5 | 0xb5 = 1.011.0101 = -0b1.0101×2 ⁻¹ = -0.65625 | 0xf5 = 1.111.0101 = -0b1.0101×2 ³ = -10.5 |
| 0x36 = 0.011.0110 = +0b1.0110×2 ⁻¹ = 0.6875 | 0x76 = 0.111.0110 = +0b1.0110×2 ³ = 11.0 | 0xb6 = 1.011.0110 = -0b1.0110×2 ⁻¹ = -0.6875 | 0xf6 = 1.111.0110 = -0b1.0110×2 ³ = -11.0 |
| 0x37 = 0.011.0111 = +0b1.0111×2 ⁻¹ = 0.71875 | 0x77 = 0.111.0111 = +0b1.0111×2 ³ = 11.5 | 0xb7 = 1.011.0111 = -0b1.0111×2 ⁻¹ = -0.71875 | 0xf7 = 1.111.0111 = -0b1.0111×2 ³ = -11.5 |
| 0x38 = 0.011.1000 = +0b1.1000×2 ⁻¹ = 0.75 | 0x78 = 0.111.1000 = +0b1.1000×2 ³ = 12.0 | 0xb8 = 1.011.1000 = -0b1.1000×2 ⁻¹ = -0.75 | 0xf8 = 1.111.1000 = -0b1.1000×2 ³ = -12.0 |
| 0x39 = 0.011.1001 = +0b1.1001×2 ⁻¹ = 0.78125 | 0x79 = 0.111.1001 = +0b1.1001×2 ³ = 12.5 | 0xb9 = 1.011.1001 = -0b1.1001×2 ⁻¹ = -0.78125 | 0xf9 = 1.111.1001 = -0b1.1001×2 ³ = -12.5 |
| 0x3a = 0.011.1010 = +0b1.1010×2 ⁻¹ = 0.8125 | 0x7a = 0.111.1010 = +0b1.1010×2 ³ = 13.0 | 0xba = 1.011.1010 = -0b1.1010×2 ⁻¹ = -0.8125 | 0xfa = 1.111.1010 = -0b1.1010×2 ³ = -13.0 |
| 0x3b = 0.011.1011 = +0b1.1011×2 ⁻¹ = 0.84375 | 0x7b = 0.111.1011 = +0b1.1011×2 ³ = 13.5 | 0xbb = 1.011.1011 = -0b1.1011×2 ⁻¹ = -0.84375 | 0xfb = 1.111.1011 = -0b1.1011×2 ³ = -13.5 |
| 0x3c = 0.011.1100 = +0b1.1100×2 ⁻¹ = 0.875 | 0x7c = 0.111.1100 = +0b1.1100×2 ³ = 14.0 | 0xbc = 1.011.1100 = -0b1.1100×2 ⁻¹ = -0.875 | 0xfc = 1.111.1100 = -0b1.1100×2 ³ = -14.0 |
| 0x3d = 0.011.1101 = +0b1.1101×2 ⁻¹ = 0.90625 | 0x7d = 0.111.1101 = +0b1.1101×2 ³ = 14.5 | 0xbd = 1.011.1101 = -0b1.1101×2 ⁻¹ = -0.90625 | 0xfd = 1.111.1101 = -0b1.1101×2 ³ = -14.5 |
| 0x3e = 0.011.1110 = +0b1.1110×2 ⁻¹ = 0.9375 | 0x7e = 0.111.1110 = +0b1.1110×2 ³ = 15.0 | 0xbe = 1.011.1110 = -0b1.1110×2 ⁻¹ = -0.9375 | 0xfe = 1.111.1110 = -0b1.1110×2 ³ = -15.0 |
| 0x3f = 0.011.1111 = +0b1.1111×2 ⁻¹ = 0.96875 | 0x7f = 0.111.1111 = +Inf | 0xbf = 1.011.1111 = -0b1.1111×2 ⁻¹ = -0.96875 | 0xff = 1.111.1111 = -Inf |

C.5 Value Table: P6, $p = 6$, $\text{emax} = 1$

| | | | |
|---|---|--|--|
| 0x00 = 0.00.00000 = 0.0 | 0x40 = 0.10.00000 = +0b1.00000×2 ⁰ = 1.0 | 0x80 = 1.00.00000 = NaN | 0xc0 = 1.10.00000 = -0b1.00000×2 ⁰ = -1.0 |
| 0x01 = 0.00.00001 = +0b0.00001×2 ⁻¹ = 0.015625 | 0x41 = 0.10.00001 = +0b1.00001×2 ⁰ = 1.03125 | 0x81 = 1.00.00001 = -0b0.00001×2 ⁻¹ = -0.015625 | 0xc1 = 1.10.00001 = -0b1.00001×2 ⁰ = -1.03125 |
| 0x02 = 0.00.00010 = +0b0.00010×2 ⁻¹ = 0.03125 | 0x42 = 0.10.00010 = +0b1.00010×2 ⁰ = 1.0625 | 0x82 = 1.00.00010 = -0b0.00010×2 ⁻¹ = -0.03125 | 0xc2 = 1.10.00010 = -0b1.00010×2 ⁰ = -1.0625 |
| 0x03 = 0.00.00011 = +0b0.00011×2 ⁻¹ = 0.046875 | 0x43 = 0.10.00011 = +0b1.00011×2 ⁰ = 1.09375 | 0x83 = 1.00.00011 = -0b0.00011×2 ⁻¹ = -0.046875 | 0xc3 = 1.10.00011 = -0b1.00011×2 ⁰ = -1.09375 |
| 0x04 = 0.00.00100 = +0b0.00100×2 ⁻¹ = 0.0625 | 0x44 = 0.10.00100 = +0b1.00100×2 ⁰ = 1.125 | 0x84 = 1.00.00100 = -0b0.00100×2 ⁻¹ = -0.0625 | 0xc4 = 1.10.00100 = -0b1.00100×2 ⁰ = -1.125 |
| 0x05 = 0.00.00101 = +0b0.00101×2 ⁻¹ = 0.078125 | 0x45 = 0.10.00101 = +0b1.00101×2 ⁰ = 1.15625 | 0x85 = 1.00.00101 = -0b0.00101×2 ⁻¹ = -0.078125 | 0xc5 = 1.10.00101 = -0b1.00101×2 ⁰ = -1.15625 |
| 0x06 = 0.00.00110 = +0b0.00110×2 ⁻¹ = 0.09375 | 0x46 = 0.10.00110 = +0b1.00110×2 ⁰ = 1.1875 | 0x86 = 1.00.00110 = -0b0.00110×2 ⁻¹ = -0.09375 | 0xc6 = 1.10.00110 = -0b1.00110×2 ⁰ = -1.1875 |
| 0x07 = 0.00.00111 = +0b0.00111×2 ⁻¹ = 0.109375 | 0x47 = 0.10.00111 = +0b1.00111×2 ⁰ = 1.21875 | 0x87 = 1.00.00111 = -0b0.00111×2 ⁻¹ = -0.109375 | 0xc7 = 1.10.00111 = -0b1.00111×2 ⁰ = -1.21875 |
| 0x08 = 0.00.01000 = +0b0.01000×2 ⁻¹ = 0.125 | 0x48 = 0.10.01000 = +0b1.01000×2 ⁰ = 1.25 | 0x88 = 1.00.01000 = -0b0.01000×2 ⁻¹ = -0.125 | 0xc8 = 1.10.01000 = -0b1.01000×2 ⁰ = -1.25 |
| 0x09 = 0.00.01001 = +0b0.01001×2 ⁻¹ = 0.140625 | 0x49 = 0.10.01001 = +0b1.01001×2 ⁰ = 1.28125 | 0x89 = 1.00.01001 = -0b0.01001×2 ⁻¹ = -0.140625 | 0xc9 = 1.10.01001 = -0b1.01001×2 ⁰ = -1.28125 |
| 0x0a = 0.00.01010 = +0b0.01010×2 ⁻¹ = 0.15625 | 0x4a = 0.10.01010 = +0b1.01010×2 ⁰ = 1.3125 | 0x8a = 1.00.01010 = -0b0.01010×2 ⁻¹ = -0.15625 | 0xca = 1.10.01010 = -0b1.01010×2 ⁰ = -1.3125 |
| 0x0b = 0.00.01011 = +0b0.01011×2 ⁻¹ = 0.171875 | 0x4b = 0.10.01011 = +0b1.01011×2 ⁰ = 1.34375 | 0x8b = 1.00.01011 = -0b0.01011×2 ⁻¹ = -0.171875 | 0xcb = 1.10.01011 = -0b1.01011×2 ⁰ = -1.34375 |
| 0x0c = 0.00.01100 = +0b0.01100×2 ⁻¹ = 0.1875 | 0x4c = 0.10.01100 = +0b1.01100×2 ⁰ = 1.375 | 0x8c = 1.00.01100 = -0b0.01100×2 ⁻¹ = -0.1875 | 0xcc = 1.10.01100 = -0b1.01100×2 ⁰ = -1.375 |
| 0x0d = 0.00.01101 = +0b0.01101×2 ⁻¹ = 0.203125 | 0x4d = 0.10.01101 = +0b1.01101×2 ⁰ = 1.40625 | 0x8d = 1.00.01101 = -0b0.01101×2 ⁻¹ = -0.203125 | 0xcd = 1.10.01101 = -0b1.01101×2 ⁰ = -1.40625 |
| 0x0e = 0.00.01110 = +0b0.01110×2 ⁻¹ = 0.21875 | 0x4e = 0.10.01110 = +0b1.01110×2 ⁰ = 1.4375 | 0x8e = 1.00.01110 = -0b0.01110×2 ⁻¹ = -0.21875 | 0xce = 1.10.01110 = -0b1.01110×2 ⁰ = -1.4375 |
| 0x0f = 0.00.01111 = +0b0.01111×2 ⁻¹ = 0.234375 | 0x4f = 0.10.01111 = +0b1.01111×2 ⁰ = 1.46875 | 0x8f = 1.00.01111 = -0b0.01111×2 ⁻¹ = -0.234375 | 0xcf = 1.10.01111 = -0b1.01111×2 ⁰ = -1.46875 |
| 0x10 = 0.00.10000 = +0b0.10000×2 ⁻¹ = 0.25 | 0x50 = 0.10.10000 = +0b1.10000×2 ⁰ = 1.5 | 0x90 = 1.00.10000 = -0b0.10000×2 ⁻¹ = -0.25 | 0xd0 = 1.10.10000 = -0b1.10000×2 ⁰ = -1.5 |
| 0x11 = 0.00.10001 = +0b0.10001×2 ⁻¹ = 0.265625 | 0x51 = 0.10.10001 = +0b1.10001×2 ⁰ = 1.53125 | 0x91 = 1.00.10001 = -0b0.10001×2 ⁻¹ = -0.265625 | 0xd1 = 1.10.10001 = -0b1.10001×2 ⁰ = -1.53125 |
| 0x12 = 0.00.10010 = +0b0.10010×2 ⁻¹ = 0.28125 | 0x52 = 0.10.10010 = +0b1.10010×2 ⁰ = 1.5625 | 0x92 = 1.00.10010 = -0b0.10010×2 ⁻¹ = -0.28125 | 0xd2 = 1.10.10010 = -0b1.10010×2 ⁰ = -1.5625 |
| 0x13 = 0.00.10011 = +0b0.10011×2 ⁻¹ = 0.296875 | 0x53 = 0.10.10011 = +0b1.10011×2 ⁰ = 1.59375 | 0x93 = 1.00.10011 = -0b0.10011×2 ⁻¹ = -0.296875 | 0xd3 = 1.10.10011 = -0b1.10011×2 ⁰ = -1.59375 |
| 0x14 = 0.00.10100 = +0b0.10100×2 ⁻¹ = 0.3125 | 0x54 = 0.10.10100 = +0b1.10100×2 ⁰ = 1.625 | 0x94 = 1.00.10100 = -0b0.10100×2 ⁻¹ = -0.3125 | 0xd4 = 1.10.10100 = -0b1.10100×2 ⁰ = -1.625 |
| 0x15 = 0.00.10101 = +0b0.10101×2 ⁻¹ = 0.328125 | 0x55 = 0.10.10101 = +0b1.10101×2 ⁰ = 1.65625 | 0x95 = 1.00.10101 = -0b0.10101×2 ⁻¹ = -0.328125 | 0xd5 = 1.10.10101 = -0b1.10101×2 ⁰ = -1.65625 |
| 0x16 = 0.00.10110 = +0b0.10110×2 ⁻¹ = 0.34375 | 0x56 = 0.10.10110 = +0b1.10110×2 ⁰ = 1.6875 | 0x96 = 1.00.10110 = -0b0.10110×2 ⁻¹ = -0.34375 | 0xd6 = 1.10.10110 = -0b1.10110×2 ⁰ = -1.6875 |
| 0x17 = 0.00.10111 = +0b0.10111×2 ⁻¹ = 0.359375 | 0x57 = 0.10.10111 = +0b1.10111×2 ⁰ = 1.71875 | 0x97 = 1.00.10111 = -0b0.10111×2 ⁻¹ = -0.359375 | 0xd7 = 1.10.10111 = -0b1.10111×2 ⁰ = -1.71875 |
| 0x18 = 0.00.11000 = +0b0.11000×2 ⁻¹ = 0.375 | 0x58 = 0.10.11000 = +0b1.11000×2 ⁰ = 1.75 | 0x98 = 1.00.11000 = -0b0.11000×2 ⁻¹ = -0.375 | 0xd8 = 1.10.11000 = -0b1.11000×2 ⁰ = -1.75 |
| 0x19 = 0.00.11001 = +0b0.11001×2 ⁻¹ = 0.390625 | 0x59 = 0.10.11001 = +0b1.11001×2 ⁰ = 1.78125 | 0x99 = 1.00.11001 = -0b0.11001×2 ⁻¹ = -0.390625 | 0xd9 = 1.10.11001 = -0b1.11001×2 ⁰ = -1.78125 |
| 0x1a = 0.00.11010 = +0b0.11010×2 ⁻¹ = 0.40625 | 0x5a = 0.10.11010 = +0b1.11010×2 ⁰ = 1.8125 | 0x9a = 1.00.11010 = -0b0.11010×2 ⁻¹ = -0.40625 | 0xda = 1.10.11010 = -0b1.11010×2 ⁰ = -1.8125 |
| 0x1b = 0.00.11011 = +0b0.11011×2 ⁻¹ = 0.421875 | 0x5b = 0.10.11011 = +0b1.11011×2 ⁰ = 1.84375 | 0x9b = 1.00.11011 = -0b0.11011×2 ⁻¹ = -0.421875 | 0xdb = 1.10.11011 = -0b1.11011×2 ⁰ = -1.84375 |
| 0x1c = 0.00.11100 = +0b0.11100×2 ⁻¹ = 0.4375 | 0x5c = 0.10.11100 = +0b1.11100×2 ⁰ = 1.875 | 0x9c = 1.00.11100 = -0b0.11100×2 ⁻¹ = -0.4375 | 0xdc = 1.10.11100 = -0b1.11100×2 ⁰ = -1.875 |
| 0x1d = 0.00.11101 = +0b0.11101×2 ⁻¹ = 0.453125 | 0x5d = 0.10.11101 = +0b1.11101×2 ⁰ = 1.90625 | 0x9d = 1.00.11101 = -0b0.11101×2 ⁻¹ = -0.453125 | 0xdd = 1.10.11101 = -0b1.11101×2 ⁰ = -1.90625 |
| 0x1e = 0.00.11110 = +0b0.11110×2 ⁻¹ = 0.46875 | 0x5e = 0.10.11110 = +0b1.11110×2 ⁰ = 1.9375 | 0x9e = 1.00.11110 = -0b0.11110×2 ⁻¹ = -0.46875 | 0xde = 1.10.11110 = -0b1.11110×2 ⁰ = -1.9375 |
| 0x1f = 0.00.11111 = +0b0.11111×2 ⁻¹ = 0.484375 | 0x5f = 0.10.11111 = +0b1.11111×2 ⁰ = 1.96875 | 0x9f = 1.00.11111 = -0b0.11111×2 ⁻¹ = -0.484375 | 0xdf = 1.10.11111 = -0b1.11111×2 ⁰ = -1.96875 |
| 0x20 = 0.01.00000 = +0b1.00000×2 ⁻¹ = 0.5 | 0x60 = 0.11.00000 = +0b1.00000×2 ¹ = 2.0 | 0xa0 = 1.01.00000 = -0b1.00000×2 ⁻¹ = -0.5 | 0xe0 = 1.11.00000 = -0b1.00000×2 ¹ = -2.0 |
| 0x21 = 0.01.00001 = +0b1.00001×2 ⁻¹ = 0.515625 | 0x61 = 0.11.00001 = +0b1.00001×2 ¹ = 2.0625 | 0xa1 = 1.01.00001 = -0b1.00001×2 ⁻¹ = -0.515625 | 0xe1 = 1.11.00001 = -0b1.00001×2 ¹ = -2.0625 |
| 0x22 = 0.01.00010 = +0b1.00010×2 ⁻¹ = 0.53125 | 0x62 = 0.11.00010 = +0b1.00010×2 ¹ = 2.125 | 0xa2 = 1.01.00010 = -0b1.00010×2 ⁻¹ = -0.53125 | 0xe2 = 1.11.00010 = -0b1.00010×2 ¹ = -2.125 |
| 0x23 = 0.01.00011 = +0b1.00011×2 ⁻¹ = 0.546875 | 0x63 = 0.11.00011 = +0b1.00011×2 ¹ = 2.1875 | 0xa3 = 1.01.00011 = -0b1.00011×2 ⁻¹ = -0.546875 | 0xe3 = 1.11.00011 = -0b1.00011×2 ¹ = -2.1875 |
| 0x24 = 0.01.00100 = +0b1.00100×2 ⁻¹ = 0.5625 | 0x64 = 0.11.00100 = +0b1.00100×2 ¹ = 2.25 | 0xa4 = 1.01.00100 = -0b1.00100×2 ⁻¹ = -0.5625 | 0xe4 = 1.11.00100 = -0b1.00100×2 ¹ = -2.25 |
| 0x25 = 0.01.00101 = +0b1.00101×2 ⁻¹ = 0.578125 | 0x65 = 0.11.00101 = +0b1.00101×2 ¹ = 2.3125 | 0xa5 = 1.01.00101 = -0b1.00101×2 ⁻¹ = -0.578125 | 0xe5 = 1.11.00101 = -0b1.00101×2 ¹ = -2.3125 |
| 0x26 = 0.01.00110 = +0b1.00110×2 ⁻¹ = 0.59375 | 0x66 = 0.11.00110 = +0b1.00110×2 ¹ = 2.375 | 0xa6 = 1.01.00110 = -0b1.00110×2 ⁻¹ = -0.59375 | 0xe6 = 1.11.00110 = -0b1.00110×2 ¹ = -2.375 |
| 0x27 = 0.01.00111 = +0b1.00111×2 ⁻¹ = 0.609375 | 0x67 = 0.11.00111 = +0b1.00111×2 ¹ = 2.4375 | 0xa7 = 1.01.00111 = -0b1.00111×2 ⁻¹ = -0.609375 | 0xe7 = 1.11.00111 = -0b1.00111×2 ¹ = -2.4375 |
| 0x28 = 0.01.01000 = +0b1.01000×2 ⁻¹ = 0.625 | 0x68 = 0.11.01000 = +0b1.01000×2 ¹ = 2.5 | 0xa8 = 1.01.01000 = -0b1.01000×2 ⁻¹ = -0.625 | 0xe8 = 1.11.01000 = -0b1.01000×2 ¹ = -2.5 |
| 0x29 = 0.01.01001 = +0b1.01001×2 ⁻¹ = 0.640625 | 0x69 = 0.11.01001 = +0b1.01001×2 ¹ = 2.5625 | 0xa9 = 1.01.01001 = -0b1.01001×2 ⁻¹ = -0.640625 | 0xe9 = 1.11.01001 = -0b1.01001×2 ¹ = -2.5625 |
| 0x2a = 0.01.01010 = +0b1.01010×2 ⁻¹ = 0.65625 | 0x6a = 0.11.01010 = +0b1.01010×2 ¹ = 2.625 | 0xaa = 1.01.01010 = -0b1.01010×2 ⁻¹ = -0.65625 | 0xea = 1.11.01010 = -0b1.01010×2 ¹ = -2.625 |
| 0x2b = 0.01.01011 = +0b1.01011×2 ⁻¹ = 0.671875 | 0x6b = 0.11.01011 = +0b1.01011×2 ¹ = 2.6875 | 0xab = 1.01.01011 = -0b1.01011×2 ⁻¹ = -0.671875 | 0xeb = 1.11.01011 = -0b1.01011×2 ¹ = -2.6875 |
| 0x2c = 0.01.01100 = +0b1.01100×2 ⁻¹ = 0.6875 | 0x6c = 0.11.01100 = +0b1.01100×2 ¹ = 2.75 | 0xac = 1.01.01100 = -0b1.01100×2 ⁻¹ = -0.6875 | 0xec = 1.11.01100 = -0b1.01100×2 ¹ = -2.75 |
| 0x2d = 0.01.01101 = +0b1.01101×2 ⁻¹ = 0.703125 | 0x6d = 0.11.01101 = +0b1.01101×2 ¹ = 2.8125 | 0xad = 1.01.01101 = -0b1.01101×2 ⁻¹ = -0.703125 | 0xed = 1.11.01101 = -0b1.01101×2 ¹ = -2.8125 |
| 0x2e = 0.01.01110 = +0b1.01110×2 ⁻¹ = 0.71875 | 0x6e = 0.11.01110 = +0b1.01110×2 ¹ = 2.875 | 0xae = 1.01.01110 = -0b1.01110×2 ⁻¹ = -0.71875 | 0xee = 1.11.01110 = -0b1.01110×2 ¹ = -2.875 |
| 0x2f = 0.01.01111 = +0b1.01111×2 ⁻¹ = 0.734375 | 0x6f = 0.11.01111 = +0b1.01111×2 ¹ = 2.9375 | 0xaf = 1.01.01111 = -0b1.01111×2 ⁻¹ = -0.734375 | 0xef = 1.11.01111 = -0b1.01111×2 ¹ = -2.9375 |
| 0x30 = 0.01.10000 = +0b1.10000×2 ⁻¹ = 0.75 | 0x70 = 0.11.10000 = +0b1.10000×2 ¹ = 3.0 | 0xb0 = 1.01.10000 = -0b1.10000×2 ⁻¹ = -0.75 | 0xf0 = 1.11.10000 = -0b1.10000×2 ¹ = -3.0 |
| 0x31 = 0.01.10001 = +0b1.10001×2 ⁻¹ = 0.765625 | 0x71 = 0.11.10001 = +0b1.10001×2 ¹ = 3.0625 | 0xb1 = 1.01.10001 = -0b1.10001×2 ⁻¹ = -0.765625 | 0xf1 = 1.11.10001 = -0b1.10001×2 ¹ = -3.0625 |
| 0x32 = 0.01.10010 = +0b1.10010×2 ⁻¹ = 0.78125 | 0x72 = 0.11.10010 = +0b1.10010×2 ¹ = 3.125 | 0xb2 = 1.01.10010 = -0b1.10010×2 ⁻¹ = -0.78125 | 0xf2 = 1.11.10010 = -0b1.10010×2 ¹ = -3.125 |
| 0x33 = 0.01.10011 = +0b1.10011×2 ⁻¹ = 0.796875 | 0x73 = 0.11.10011 = +0b1.10011×2 ¹ = 3.1875 | 0xb3 = 1.01.10011 = -0b1.10011×2 ⁻¹ = -0.796875 | 0xf3 = 1.11.10011 = -0b1.10011×2 ¹ = -3.1875 |
| 0x34 = 0.01.10100 = +0b1.10100×2 ⁻¹ = 0.8125 | 0x74 = 0.11.10100 = +0b1.10100×2 ¹ = 3.25 | 0xb4 = 1.01.10100 = -0b1.10100×2 ⁻¹ = -0.8125 | 0xf4 = 1.11.10100 = -0b1.10100×2 ¹ = -3.25 |
| 0x35 = 0.01.10101 = +0b1.10101×2 ⁻¹ = 0.828125 | 0x75 = 0.11.10101 = +0b1.10101×2 ¹ = 3.3125 | 0xb5 = 1.01.10101 = -0b1.10101×2 ⁻¹ = -0.828125 | 0xf5 = 1.11.10101 = -0b1.10101×2 ¹ = -3.3125 |
| 0x36 = 0.01.10110 = +0b1.10110×2 ⁻¹ = 0.84375 | 0x76 = 0.11.10110 = +0b1.10110×2 ¹ = 3.375 | 0xb6 = 1.01.10110 = -0b1.10110×2 ⁻¹ = -0.84375 | 0xf6 = 1.11.10110 = -0b1.10110×2 ¹ = -3.375 |
| 0x37 = 0.01.10111 = +0b1.10111×2 ⁻¹ = 0.859375 | 0x77 = 0.11.10111 = +0b1.10111×2 ¹ = 3.4375 | 0xb7 = 1.01.10111 = -0b1.10111×2 ⁻¹ = -0.859375 | 0xf7 = 1.11.10111 = -0b1.10111×2 ¹ = -3.4375 |
| 0x38 = 0.01.11000 = +0b1.11000×2 ⁻¹ = 0.875 | 0x78 = 0.11.11000 = +0b1.11000×2 ¹ = 3.5 | 0xb8 = 1.01.11000 = -0b1.11000×2 ⁻¹ = -0.875 | 0xf8 = 1.11.11000 = -0b1.11000×2 ¹ = -3.5 |
| 0x39 = 0.01.11001 = +0b1.11001×2 ⁻¹ = 0.890625 | 0x79 = 0.11.11001 = +0b1.11001×2 ¹ = 3.5625 | 0xb9 = 1.01.11001 = -0b1.11001×2 ⁻¹ = -0.890625 | 0xf9 = 1.11.11001 = -0b1.11001×2 ¹ = -3.5625 |
| 0x3a = 0.01.11010 = +0b1.11010×2 ⁻¹ = 0.90625 | 0x7a = 0.11.11010 = +0b1.11010×2 ¹ = 3.625 | 0xba = 1.01.11010 = -0b1.11010×2 ⁻¹ = -0.90625 | 0xfa = 1.11.11010 = -0b1.11010×2 ¹ = -3.625 |
| 0x3b = 0.01.11011 = +0b1.11011×2 ⁻¹ = 0.921875 | 0x7b = 0.11.11011 = +0b1.11011×2 ¹ = 3.6875 | 0xbb = 1.01.11011 = -0b1.11011×2 ⁻¹ = -0.921875 | 0xfb = 1.11.11011 = -0b1.11011×2 ¹ = -3.6875 |
| 0x3c = 0.01.11100 = +0b1.11100×2 ⁻¹ = 0.9375 | 0x7c = 0.11.11100 = +0b1.11100×2 ¹ = 3.75 | 0xbc = 1.01.11100 = -0b1.11100×2 ⁻¹ = -0.9375 | 0xfc = 1.11.11100 = -0b1.11100×2 ¹ = -3.75 |
| 0x3d = 0.01.11101 = +0b1.11101×2 ⁻¹ = 0.953125 | 0x7d = 0.11.11101 = +0b1.11101×2 ¹ = 3.8125 | 0xbd = 1.01.11101 = -0b1.11101×2 ⁻¹ = -0.953125 | 0xfd = 1.11.11101 = -0b1.11101×2 ¹ |

C.6 Value Table: P7 (Linear), $p = 7$, $\text{emax} = 0$

| | | | |
|---|---|--|--|
| 0x00 = 0.0.000000 = 0.0 | 0x40 = 0.1.000000 = +0b1.000000×2 ⁰ = 1.0 | 0x80 = 1.0.000000 = NaN | 0xc0 = 1.1.000000 = -0b1.000000×2 ⁰ = -1.0 |
| 0x01 = 0.0.000001 = +0b0.000001×2 ⁰ = 0.015625 | 0x41 = 0.1.000001 = +0b1.000001×2 ⁰ = 1.015625 | 0x81 = 1.0.000001 = -0b0.000001×2 ⁰ = -0.015625 | 0xc1 = 1.1.000001 = -0b1.000001×2 ⁰ = -1.015625 |
| 0x02 = 0.0.000010 = +0b0.000010×2 ⁰ = 0.03125 | 0x42 = 0.1.000010 = +0b1.000010×2 ⁰ = 1.03125 | 0x82 = 1.0.000010 = -0b0.000010×2 ⁰ = -0.03125 | 0xc2 = 1.1.000010 = -0b1.000010×2 ⁰ = -1.03125 |
| 0x03 = 0.0.000011 = +0b0.000011×2 ⁰ = 0.046875 | 0x43 = 0.1.000011 = +0b1.000011×2 ⁰ = 1.046875 | 0x83 = 1.0.000011 = -0b0.000011×2 ⁰ = -0.046875 | 0xc3 = 1.1.000011 = -0b1.000011×2 ⁰ = -1.046875 |
| 0x04 = 0.0.000100 = +0b0.000100×2 ⁰ = 0.0625 | 0x44 = 0.1.000100 = +0b1.000100×2 ⁰ = 1.0625 | 0x84 = 1.0.000100 = -0b0.000100×2 ⁰ = -0.0625 | 0xc4 = 1.1.000100 = -0b1.000100×2 ⁰ = -1.0625 |
| 0x05 = 0.0.000101 = +0b0.000101×2 ⁰ = 0.078125 | 0x45 = 0.1.000101 = +0b1.000101×2 ⁰ = 1.078125 | 0x85 = 1.0.000101 = -0b0.000101×2 ⁰ = -0.078125 | 0xc5 = 1.1.000101 = -0b1.000101×2 ⁰ = -1.078125 |
| 0x06 = 0.0.000110 = +0b0.000110×2 ⁰ = 0.09375 | 0x46 = 0.1.000110 = +0b1.000110×2 ⁰ = 1.09375 | 0x86 = 1.0.000110 = -0b0.000110×2 ⁰ = -0.09375 | 0xc6 = 1.1.000110 = -0b1.000110×2 ⁰ = -1.09375 |
| 0x07 = 0.0.000111 = +0b0.000111×2 ⁰ = 0.109375 | 0x47 = 0.1.000111 = +0b1.000111×2 ⁰ = 1.109375 | 0x87 = 1.0.000111 = -0b0.000111×2 ⁰ = -0.109375 | 0xc7 = 1.1.000111 = -0b1.000111×2 ⁰ = -1.109375 |
| 0x08 = 0.0.001000 = +0b0.001000×2 ⁰ = 0.125 | 0x48 = 0.1.001000 = +0b1.001000×2 ⁰ = 1.125 | 0x88 = 1.0.001000 = -0b0.001000×2 ⁰ = -0.125 | 0xc8 = 1.1.001000 = -0b1.001000×2 ⁰ = -1.125 |
| 0x09 = 0.0.001001 = +0b0.001001×2 ⁰ = 0.140625 | 0x49 = 0.1.001001 = +0b1.001001×2 ⁰ = 1.140625 | 0x89 = 1.0.001001 = -0b0.001001×2 ⁰ = -0.140625 | 0xc9 = 1.1.001001 = -0b1.001001×2 ⁰ = -1.140625 |
| 0x0a = 0.0.001010 = +0b0.001010×2 ⁰ = 0.15625 | 0x4a = 0.1.001010 = +0b1.001010×2 ⁰ = 1.15625 | 0x8a = 1.0.001010 = -0b0.001010×2 ⁰ = -0.15625 | 0xca = 1.1.001010 = -0b1.001010×2 ⁰ = -1.15625 |
| 0x0b = 0.0.001011 = +0b0.001011×2 ⁰ = 0.171875 | 0x4b = 0.1.001011 = +0b1.001011×2 ⁰ = 1.171875 | 0x8b = 1.0.001011 = -0b0.001011×2 ⁰ = -0.171875 | 0xcb = 1.1.001011 = -0b1.001011×2 ⁰ = -1.171875 |
| 0x0c = 0.0.001100 = +0b0.001100×2 ⁰ = 0.1875 | 0x4c = 0.1.001100 = +0b1.001100×2 ⁰ = 1.1875 | 0x8c = 1.0.001100 = -0b0.001100×2 ⁰ = -0.1875 | 0xcc = 1.1.001100 = -0b1.001100×2 ⁰ = -1.1875 |
| 0x0d = 0.0.001101 = +0b0.001101×2 ⁰ = 0.203125 | 0x4d = 0.1.001101 = +0b1.001101×2 ⁰ = 1.203125 | 0x8d = 1.0.001101 = -0b0.001101×2 ⁰ = -0.203125 | 0xcd = 1.1.001101 = -0b1.001101×2 ⁰ = -1.203125 |
| 0x0e = 0.0.001110 = +0b0.001110×2 ⁰ = 0.21875 | 0x4e = 0.1.001110 = +0b1.001110×2 ⁰ = 1.21875 | 0x8e = 1.0.001110 = -0b0.001110×2 ⁰ = -0.21875 | 0xce = 1.1.001110 = -0b1.001110×2 ⁰ = -1.21875 |
| 0x0f = 0.0.001111 = +0b0.001111×2 ⁰ = 0.234375 | 0x4f = 0.1.001111 = +0b1.001111×2 ⁰ = 1.234375 | 0x8f = 1.0.001111 = -0b0.001111×2 ⁰ = -0.234375 | 0xcf = 1.1.001111 = -0b1.001111×2 ⁰ = -1.234375 |
| 0x10 = 0.0.010000 = +0b0.010000×2 ⁰ = 0.25 | 0x50 = 0.1.010000 = +0b1.010000×2 ⁰ = 1.25 | 0x90 = 1.0.010000 = -0b0.010000×2 ⁰ = -0.25 | 0xd0 = 1.1.010000 = -0b1.010000×2 ⁰ = -1.25 |
| 0x11 = 0.0.010001 = +0b0.010001×2 ⁰ = 0.265625 | 0x51 = 0.1.010001 = +0b1.010001×2 ⁰ = 1.265625 | 0x91 = 1.0.010001 = -0b0.010001×2 ⁰ = -0.265625 | 0xd1 = 1.1.010001 = -0b1.010001×2 ⁰ = -1.265625 |
| 0x12 = 0.0.010010 = +0b0.010010×2 ⁰ = 0.28125 | 0x52 = 0.1.010010 = +0b1.010010×2 ⁰ = 1.28125 | 0x92 = 1.0.010010 = -0b0.010010×2 ⁰ = -0.28125 | 0xd2 = 1.1.010010 = -0b1.010010×2 ⁰ = -1.28125 |
| 0x13 = 0.0.010011 = +0b0.010011×2 ⁰ = 0.296875 | 0x53 = 0.1.010011 = +0b1.010011×2 ⁰ = 1.296875 | 0x93 = 1.0.010011 = -0b0.010011×2 ⁰ = -0.296875 | 0xd3 = 1.1.010011 = -0b1.010011×2 ⁰ = -1.296875 |
| 0x14 = 0.0.010100 = +0b0.010100×2 ⁰ = 0.3125 | 0x54 = 0.1.010100 = +0b1.010100×2 ⁰ = 1.3125 | 0x94 = 1.0.010100 = -0b0.010100×2 ⁰ = -0.3125 | 0xd4 = 1.1.010100 = -0b1.010100×2 ⁰ = -1.3125 |
| 0x15 = 0.0.010101 = +0b0.010101×2 ⁰ = 0.328125 | 0x55 = 0.1.010101 = +0b1.010101×2 ⁰ = 1.328125 | 0x95 = 1.0.010101 = -0b0.010101×2 ⁰ = -0.328125 | 0xd5 = 1.1.010101 = -0b1.010101×2 ⁰ = -1.328125 |
| 0x16 = 0.0.010110 = +0b0.010110×2 ⁰ = 0.34375 | 0x56 = 0.1.010110 = +0b1.010110×2 ⁰ = 1.34375 | 0x96 = 1.0.010110 = -0b0.010110×2 ⁰ = -0.34375 | 0xd6 = 1.1.010110 = -0b1.010110×2 ⁰ = -1.34375 |
| 0x17 = 0.0.010111 = +0b0.010111×2 ⁰ = 0.359375 | 0x57 = 0.1.010111 = +0b1.010111×2 ⁰ = 1.359375 | 0x97 = 1.0.010111 = -0b0.010111×2 ⁰ = -0.359375 | 0xd7 = 1.1.010111 = -0b1.010111×2 ⁰ = -1.359375 |
| 0x18 = 0.0.011000 = +0b0.011000×2 ⁰ = 0.375 | 0x58 = 0.1.011000 = +0b1.011000×2 ⁰ = 1.375 | 0x98 = 1.0.011000 = -0b0.011000×2 ⁰ = -0.375 | 0xd8 = 1.1.011000 = -0b1.011000×2 ⁰ = -1.375 |
| 0x19 = 0.0.011001 = +0b0.011001×2 ⁰ = 0.390625 | 0x59 = 0.1.011001 = +0b1.011001×2 ⁰ = 1.390625 | 0x99 = 1.0.011001 = -0b0.011001×2 ⁰ = -0.390625 | 0xd9 = 1.1.011001 = -0b1.011001×2 ⁰ = -1.390625 |
| 0x1a = 0.0.011010 = +0b0.011010×2 ⁰ = 0.40625 | 0x5a = 0.1.011010 = +0b1.011010×2 ⁰ = 1.40625 | 0x9a = 1.0.011010 = -0b0.011010×2 ⁰ = -0.40625 | 0xda = 1.1.011010 = -0b1.011010×2 ⁰ = -1.40625 |
| 0x1b = 0.0.011011 = +0b0.011011×2 ⁰ = 0.421875 | 0x5b = 0.1.011011 = +0b1.011011×2 ⁰ = 1.421875 | 0x9b = 1.0.011011 = -0b0.011011×2 ⁰ = -0.421875 | 0xdb = 1.1.011011 = -0b1.011011×2 ⁰ = -1.421875 |
| 0x1c = 0.0.011100 = +0b0.011100×2 ⁰ = 0.4375 | 0x5c = 0.1.011100 = +0b1.011100×2 ⁰ = 1.4375 | 0x9c = 1.0.011100 = -0b0.011100×2 ⁰ = -0.4375 | 0xdc = 1.1.011100 = -0b1.011100×2 ⁰ = -1.4375 |
| 0x1d = 0.0.011101 = +0b0.011101×2 ⁰ = 0.453125 | 0x5d = 0.1.011101 = +0b1.011101×2 ⁰ = 1.453125 | 0x9d = 1.0.011101 = -0b0.011101×2 ⁰ = -0.453125 | 0xdd = 1.1.011101 = -0b1.011101×2 ⁰ = -1.453125 |
| 0x1e = 0.0.011110 = +0b0.011110×2 ⁰ = 0.46875 | 0x5e = 0.1.011110 = +0b1.011110×2 ⁰ = 1.46875 | 0x9e = 1.0.011110 = -0b0.011110×2 ⁰ = -0.46875 | 0xde = 1.1.011110 = -0b1.011110×2 ⁰ = -1.46875 |
| 0x1f = 0.0.011111 = +0b0.011111×2 ⁰ = 0.484375 | 0x5f = 0.1.011111 = +0b1.011111×2 ⁰ = 1.484375 | 0x9f = 1.0.011111 = -0b0.011111×2 ⁰ = -0.484375 | 0xdf = 1.1.011111 = -0b1.011111×2 ⁰ = -1.484375 |
| 0x20 = 0.0.100000 = +0b0.100000×2 ⁰ = 0.5 | 0x60 = 0.1.100000 = +0b1.100000×2 ⁰ = 1.5 | 0xa0 = 1.0.100000 = -0b0.100000×2 ⁰ = -0.5 | 0xe0 = 1.1.100000 = -0b1.100000×2 ⁰ = -1.5 |
| 0x21 = 0.0.100001 = +0b0.100001×2 ⁰ = 0.515625 | 0x61 = 0.1.100001 = +0b1.100001×2 ⁰ = 1.515625 | 0xa1 = 1.0.100001 = -0b0.100001×2 ⁰ = -0.515625 | 0xe1 = 1.1.100001 = -0b1.100001×2 ⁰ = -1.515625 |
| 0x22 = 0.0.100010 = +0b0.100010×2 ⁰ = 0.53125 | 0x62 = 0.1.100010 = +0b1.100010×2 ⁰ = 1.53125 | 0xa2 = 1.0.100010 = -0b0.100010×2 ⁰ = -0.53125 | 0xe2 = 1.1.100010 = -0b1.100010×2 ⁰ = -1.53125 |
| 0x23 = 0.0.100011 = +0b0.100011×2 ⁰ = 0.546875 | 0x63 = 0.1.100011 = +0b1.100011×2 ⁰ = 1.546875 | 0xa3 = 1.0.100011 = -0b0.100011×2 ⁰ = -0.546875 | 0xe3 = 1.1.100011 = -0b1.100011×2 ⁰ = -1.546875 |
| 0x24 = 0.0.100100 = +0b0.100100×2 ⁰ = 0.5625 | 0x64 = 0.1.100100 = +0b1.100100×2 ⁰ = 1.5625 | 0xa4 = 1.0.100100 = -0b0.100100×2 ⁰ = -0.5625 | 0xe4 = 1.1.100100 = -0b1.100100×2 ⁰ = -1.5625 |
| 0x25 = 0.0.100101 = +0b0.100101×2 ⁰ = 0.578125 | 0x65 = 0.1.100101 = +0b1.100101×2 ⁰ = 1.578125 | 0xa5 = 1.0.100101 = -0b0.100101×2 ⁰ = -0.578125 | 0xe5 = 1.1.100101 = -0b1.100101×2 ⁰ = -1.578125 |
| 0x26 = 0.0.100110 = +0b0.100110×2 ⁰ = 0.59375 | 0x66 = 0.1.100110 = +0b1.100110×2 ⁰ = 1.59375 | 0xa6 = 1.0.100110 = -0b0.100110×2 ⁰ = -0.59375 | 0xe6 = 1.1.100110 = -0b1.100110×2 ⁰ = -1.59375 |
| 0x27 = 0.0.100111 = +0b0.100111×2 ⁰ = 0.609375 | 0x67 = 0.1.100111 = +0b1.100111×2 ⁰ = 1.609375 | 0xa7 = 1.0.100111 = -0b0.100111×2 ⁰ = -0.609375 | 0xe7 = 1.1.100111 = -0b1.100111×2 ⁰ = -1.609375 |
| 0x28 = 0.0.101000 = +0b0.101000×2 ⁰ = 0.625 | 0x68 = 0.1.101000 = +0b1.101000×2 ⁰ = 1.625 | 0xa8 = 1.0.101000 = -0b0.101000×2 ⁰ = -0.625 | 0xe8 = 1.1.101000 = -0b1.101000×2 ⁰ = -1.625 |
| 0x29 = 0.0.101001 = +0b0.101001×2 ⁰ = 0.640625 | 0x69 = 0.1.101001 = +0b1.101001×2 ⁰ = 1.640625 | 0xa9 = 1.0.101001 = -0b0.101001×2 ⁰ = -0.640625 | 0xe9 = 1.1.101001 = -0b1.101001×2 ⁰ = -1.640625 |
| 0x2a = 0.0.101010 = +0b0.101010×2 ⁰ = 0.65625 | 0x6a = 0.1.101010 = +0b1.101010×2 ⁰ = 1.65625 | 0xaa = 1.0.101010 = -0b0.101010×2 ⁰ = -0.65625 | 0xea = 1.1.101010 = -0b1.101010×2 ⁰ = -1.65625 |
| 0x2b = 0.0.101011 = +0b0.101011×2 ⁰ = 0.671875 | 0x6b = 0.1.101011 = +0b1.101011×2 ⁰ = 1.671875 | 0xab = 1.0.101011 = -0b0.101011×2 ⁰ = -0.671875 | 0xeb = 1.1.101011 = -0b1.101011×2 ⁰ = -1.671875 |
| 0x2c = 0.0.101100 = +0b0.101100×2 ⁰ = 0.6875 | 0x6c = 0.1.101100 = +0b1.101100×2 ⁰ = 1.6875 | 0xac = 1.0.101100 = -0b0.101100×2 ⁰ = -0.6875 | 0xec = 1.1.101100 = -0b1.101100×2 ⁰ = -1.6875 |
| 0x2d = 0.0.101101 = +0b0.101101×2 ⁰ = 0.703125 | 0x6d = 0.1.101101 = +0b1.101101×2 ⁰ = 1.703125 | 0xad = 1.0.101101 = -0b0.101101×2 ⁰ = -0.703125 | 0xed = 1.1.101101 = -0b1.101101×2 ⁰ = -1.703125 |
| 0x2e = 0.0.101110 = +0b0.101110×2 ⁰ = 0.71875 | 0x6e = 0.1.101110 = +0b1.101110×2 ⁰ = 1.71875 | 0xae = 1.0.101110 = -0b0.101110×2 ⁰ = -0.71875 | 0xee = 1.1.101110 = -0b1.101110×2 ⁰ = -1.71875 |
| 0x2f = 0.0.101111 = +0b0.101111×2 ⁰ = 0.734375 | 0x6f = 0.1.101111 = +0b1.101111×2 ⁰ = 1.734375 | 0xaf = 1.0.101111 = -0b0.101111×2 ⁰ = -0.734375 | 0xef = 1.1.101111 = -0b1.101111×2 ⁰ = -1.734375 |
| 0x30 = 0.0.110000 = +0b0.110000×2 ⁰ = 0.75 | 0x70 = 0.1.110000 = +0b1.110000×2 ⁰ = 1.75 | 0xb0 = 1.0.110000 = -0b0.110000×2 ⁰ = -0.75 | 0xf0 = 1.1.110000 = -0b1.110000×2 ⁰ = -1.75 |
| 0x31 = 0.0.110001 = +0b0.110001×2 ⁰ = 0.765625 | 0x71 = 0.1.110001 = +0b1.110001×2 ⁰ = 1.765625 | 0xb1 = 1.0.110001 = -0b0.110001×2 ⁰ = -0.765625 | 0xf1 = 1.1.110001 = -0b1.110001×2 ⁰ = -1.765625 |
| 0x32 = 0.0.110010 = +0b0.110010×2 ⁰ = 0.78125 | 0x72 = 0.1.110010 = +0b1.110010×2 ⁰ = 1.78125 | 0xb2 = 1.0.110010 = -0b0.110010×2 ⁰ = -0.78125 | 0xf2 = 1.1.110010 = -0b1.110010×2 ⁰ = -1.78125 |
| 0x33 = 0.0.110011 = +0b0.110011×2 ⁰ = 0.796875 | 0x73 = 0.1.110011 = +0b1.110011×2 ⁰ = 1.796875 | 0xb3 = 1.0.110011 = -0b0.110011×2 ⁰ = -0.796875 | 0xf3 = 1.1.110011 = -0b1.110011×2 ⁰ = -1.796875 |
| 0x34 = 0.0.110100 = +0b0.110100×2 ⁰ = 0.8125 | 0x74 = 0.1.110100 = +0b1.110100×2 ⁰ = 1.8125 | 0xb4 = 1.0.110100 = -0b0.110100×2 ⁰ = -0.8125 | 0xf4 = 1.1.110100 = -0b1.110100×2 ⁰ = -1.8125 |
| 0x35 = 0.0.110101 = +0b0.110101×2 ⁰ = 0.828125 | 0x75 = 0.1.110101 = +0b1.110101×2 ⁰ = 1.828125 | 0xb5 = 1.0.110101 = -0b0.110101×2 ⁰ = -0.828125 | 0xf5 = 1.1.110101 = -0b1.110101×2 ⁰ = -1.828125 |
| 0x36 = 0.0.110110 = +0b0.110110×2 ⁰ = 0.84375 | 0x76 = 0.1.110110 = +0b1.110110×2 ⁰ = 1.84375 | 0xb6 = 1.0.110110 = -0b0.110110×2 ⁰ = -0.84375 | 0xf6 = 1.1.110110 = -0b1.110110×2 ⁰ = -1.84375 |
| 0x37 = 0.0.110111 = +0b0.110111×2 ⁰ = 0.859375 | 0x77 = 0.1.110111 = +0b1.110111×2 ⁰ = 1.859375 | 0xb7 = 1.0.110111 = -0b0.110111×2 ⁰ = -0.859375 | 0xf7 = 1.1.110111 = -0b1.110111×2 ⁰ = -1.859375 |
| 0x38 = 0.0.111000 = +0b0.111000×2 ⁰ = 0.875 | 0x78 = 0.1.111000 = +0b1.111000×2 ⁰ = 1.875 | 0xb8 = 1.0.111000 = -0b0.111000×2 ⁰ = -0.875 | 0xf8 = 1.1.111000 = -0b1.111000×2 ⁰ = -1.875 |
| 0x39 = 0.0.111001 = +0b0.111001×2 ⁰ = 0.890625 | 0x79 = 0.1.111001 = +0b1.111001×2 ⁰ = 1.890625 | 0xb9 = 1.0.111001 = -0b0.111001×2 ⁰ = -0.890625 | 0xf9 = 1.1.111001 = -0b1.111001×2 ⁰ = -1.890625 |
| 0x3a = 0.0.111010 = +0b0.111010×2 ⁰ = 0.90625 | 0x7a = 0.1.111010 = +0b1.111010×2 ⁰ = 1.90625 | 0xba = 1.0.111010 = -0b0.111010×2 ⁰ = -0.90625 | 0xfa = 1.1.111010 = -0b1.111010×2 ⁰ = -1.90625 |
| 0x3b = 0.0.111011 = +0b0.111011×2 ⁰ = 0.921875 | 0x7b = 0.1.111011 = +0b1.111011×2 ⁰ = 1.921875 | 0xbb = 1.0.111011 = -0b0.111011×2 ⁰ = -0.921875 | 0xfb = 1.1.111011 = -0b1. |

References

- [1] PyTorch authors. Pytorch torchtext package: `_t5_multi_head_attention_forward` . <https://github.com/pytorch/text/blob/a933cbe5a008bc2cb61d985cf5864069194157eb/torchtext/prototype/models/t5/modules.py#L236>.
- [2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter 6.2.2.3 Softmax Units for Multinoulli Output Distributions, pages 180–184. MIT Press, 2016.
- [3] Google. Jax lax package: `_float_to_int_for_sort` . https://github.com/google/jax/blob/fc5960f2b8b7a0ef74dbae4e27c5c08ff1564cff/jax/_src/lax/lax.py#L3934.
- [4] W. Kahan. Branch cuts for complex elementary functions or much ado about nothing’s sign bit. *Inst. Math. Appl. Conf. Ser. New Ser.*, 1987.
- [5] W. Kahan and J. W. Thomas. Augmenting a programming language with complex arithmetic. Technical report, EECS Department, University of California, Berkeley, 1991.
- [6] P. Micikevicius, S. Oberman, P. Dubey, M. Cornea, A. Rodriguez, I. Bratt, R. Grisenthwaite, N. Jouppi, C. Chou, A. Huffman, M. Schulte, R. Wittig, D. Jani, and S. Deng. OCP 8-bit floating point specification (OFP8). Technical report, opencompute.org, 2023.
- [7] B. Nouné, P. Jones, D. Justus, D. Masters, and C. Luschi. 8-bit numerical formats for deep neural networks. Technical report, arXiv cs.LG, 2022. <https://arxiv.org/abs/2206.02915>.
- [8] Tesla, Inc. Tesla Dojo Technology: A guide to Tesla’s configurable floating point formats and arithmetic, 2023. https://web.archive.org/web/20230503235751/https://tesla-cdn.thron.com/static/MXMU3S_tesla-dojo-technology_1WDVZN.pdf.