Index

1/n expansion, 9, 10, 31, 62, 65, 72, 93, 95, 100–103, 137, 144, 153, 162, 167, 204, 210, 211, 221, 226, 228, 229, 319, 350, 394, 396, 415

*:polation, 277, 279, 280, 411
curvature, 281
deep linear networks, 277
nonlinear networks, 278, 280
same for inter- and extra-, 281

δ expansion, 116, 269, 274
generalized to \( \epsilon_{1,2} \) expansion, 279

\( \gamma_{[0]} \) basis, 115
\( \sigma \), 120, 271
\( \sigma' \sigma' \), 271
frozen NTK, 271
kernel, 172, 269, 410
for hybrid approach, 437
output matrix, 172

absolute certainty, 156, 158, 168, 178, 260, 401
*:polation, 281
action, 11, 26, 27, 33, 72, 401, 411
effective, see effective action
quadratic, see also Gaussian distribution, 27, 66, 75, 206
quartic, see also nearly-Gaussian distribution, 28, 29–32, 34
sextic, 395
truncation, see also hierarchical scaling, 34
activation, 39
activation function, 39, 43, 244
GELU, 47, 113, 132, 135–137, 336, 346, 381
leaky ReLU, 44, 46
linear, 46, 53, 54, 111, 123, 124, 138, 140, 146, 150, 151, 233, 276, 277, 289, 346, 421
monomial, 129
perceptron, 40, 43–46, 128
quadratic, 111
ReLU, 45–47, 110, 116, 123, 125, 128, 135–137, 139, 140, 143, 148–150, 233, 238, 244, 245, 335, 336, 345, 346, 381, 395, 421, 422, 438
sigmoid, 44–46, 126, 128, 129, 244
sin, 45, 113, 131, 134, 236, 238–240, 245, 273
softplus, 46, 128, 129, 132, 135, 137
SWISH, 46, 47, 113, 126, 132, 135–137, 336, 346, 381
tanh, 45, 46, 113, 126–131, 133, 134, 141, 143, 171, 236, 238–240, 244, 245, 273, 343, 422
adventure for thrill seekers, 310, 379
algorithm dependence, 327, 348, 361, 373, 375, 382, 393
algorithm independence, 257, 258, 348, 383
algorithm projector, 327, 336, 372, 373, 375, 379, 382, 383, 393
Anderson, Thomas A. “Neo”, 21
applause, 180
architecture, 7, 40, 42, 109, 197
architecture hyperparameters, 40, 53, 64, 157
Armstrong, Neil, 252
artificial intelligence, 1, 37, 38, 39, 53
artificial neural network, see neural network
artificial neuron, 37, 39, 180, 182
atom, 3, 58
attention, see also transformer, 42

445
Index

backpropagation, 45, 189, 205, 241
backward equation
  MLP, 205, 250
backward pass, 205
Banks–Zaks fixed point, see fixed point
batch, see also stochastic gradient descent, 195, 257
Bayes, Reverend Thomas, 163
Bayesian inference, 153–155, 156, 157, 160,
  161, 163, 164, 167, 168, 191, 249,
  255, 258, 263, 264
connection to linear models, 288
evidence, 154, 156, 159, 165, 166,
  168–173, 406
relation to generalization error, 268
hierarchical modeling, 166
hypothesis, see hypothesis (Bayesian inference)
likelihood, 156, 159–161, 165, 167, 178,
  186, 187
model comparison, 154, 156, 165,
  165–170, 173, 402, 406
Bayes’ factor, 166, 167, 168, 171, 402,
  406
model fitting, 153, 156, 160, 161, 164,
  165, 167, 194, 259, 262
approximation methods, 161
exact marginalization, 162
posterior, see posterior
practicalities, 177, 185
prediction, 154, 161, 164, 165
prior, see prior
via gradient descent
  at infinite width, 263
  but not at finite width, 384
wiring
  finite width, 182
  infinite width, 179
Bayesian probability, 154, 155
  Bayes’ rule, 156, 159, 163–165, 168,
  178, 186, 191
hypothesis, see hypothesis (Bayesian inference)
  product rule, 155, 157
  statements, 154, 155
  sum rule, 155, 157
bell curve, see also Gaussian function, 12, 13
bias–variance decomposition, see also
generalization error, 267
bias–variance tradeoff, 194, 266, 269, 379, 411
for a universality class
  \( K^* = 0 \) activations, 271
  scale-invariant activations, 275
generalized, 266, 275
vs. standard, 266
relation to criticality, 269
biases, see also model parameters, 39, 193
big tech, 62
biological neural network, see also brain, 1,
  247
biological neuron, 37, 39, 154, 180
bit (unit of entropy), 401
black box, 2
blueprint, 436
Boltzmann constant, 401, 413
Boltzmann distribution, 412, 413
Boltzmann entropy, see entropy
Boltzmann, Ludwig, 3, 399
bona fide, 351
bottleneck, 55
bra-ket notation, see also Gaussian expectation, 27, 29, 30
brain, 1, 39, 42, 53
brand awareness, 375
Brown, Emmett Lathrop “Doc”, 425
Carnot, Sadi, 2
central limit theorem, 48
chain rule, 196, 201, 202, 204, 205, 241
chain-rule factor, 201, 204, 205, 214–243
channel, see also convolutional neural
  network, 42
chaos, see also overly deep, 65, 397, 426
checkpoint, 395
classification, 170, 177, 192, 260, 342, 410
CNN, see convolutional neural network
coarse-graining, see also renormalization
  group flow, see also representation
  group flow, 73, 106
Index

cognitive science, 37
cokurtosis, see kurtosis, excess
Coleman, Sidney, 191, 199
complete the square, 19, 176, 213, 214
computer vision, 42, 55, 168, 192, 428, 436
connected correlator, 23, 33, 227, 397
four-point, see also four-point vertex, see also kurtosis, excess, 24, 28, 31, 53, 59, 62, 63, 70, 182, 207
general definition, 24
higher-point, 34
odd-point vanish with parity, 23
one-point, see also mean, 23
relation to nearly-Gaussian distributions, 26
six-point, 70
two-point, see also covariance, see also metric, 23
continuum limit, see gradient descent, 360
convex function, 404
ConvNet, see convolutional neural network
convolutional neural network, 42, 55, 157, 166, 168, 428
with residual connections, see ResNet correlator
2m-point, 67
M-point, 22, 55
connected, see connected correlator
four-point, 60, 62, 208, 209
full, see also moment, 23
higher-point, 59, 65, 68
six-point, 66
two-point, 56, 57, 208, 209
coupling, 11
data-dependent, see data-dependent coupling
non-Gaussian, 32, 33
quadratic, 33, 34, 98, 100, 103, 180, 181, 183, 211
quartic, 28, 29, 31, 32, 34, 62, 70, 98, 99, 103, 180, 181, 183, 185, 211, 416
running, see running coupling
sextic, 395, 416
covariance, see also cumulant, 16, 23
critical initialization hyperparameters, see initialization hyperparameters
critical phenomena, 58
critical temperature, 59
as unsupervised learning, 422
principle of, 110, 154, 273, 276, 360
semi-criticality, 125, 132, 141
cross correlation
dNTK–preactivation, 303
P-recursion, 307
Q-recursion, 308
D-recursion, 219
F-recursion, 220
cross entropy, 404
cross-entropy loss, see loss
cubic model, see nonlinear model
cumulant, see also connected correlator, 23, 413
first (mean), see also mean, 23
general definition, 24
second (covariance), see also covariance, 23
cutoff, effective theory, 110, 144, 145, 232, 292, 312, 381, 395, 407, 408
nearly-kernel methods, 330
vs. measurement precision, 407
damping force, 366, 367, 369, 371, 375
data dependence, see also connected correlator, see also data-dependent coupling, 133, 134, 142
data distribution, see also input data, 192, 193, 194
data-dependent coupling, 93, 98, 103, 163, 167, 227, 390, 392–396
dataset, see input data
ddNTKs, 335–337, 338
  contribution to finite-width prediction, 377
full expressions, 384
scaling laws, 341
statistics, 339
  $R$-recursion, 340, 386
  $S$-recursion, 341, 387
  $T$-recursion, 341, 387
  $U$-recursion, 341, 388
step-independence, 358
deep learning, 1, 10, 39, 45, 109, 153, 156, 165, 179, 195, 205, 227, 238, 241, 389, 397
abstracted, 317
deep but not yet learning, 153
history, 38
deep linear network, see also linear, 53, 53–55, 57–60, 65, 110, 111, 113, 125, 137, 138, 140, 146, 152, 242, 276, 277, 281, 289, 397
limitations, 277
definition
  Gaussian distribution, 28, 33, 413, 414
  linear model, 318
  quadratic model, 332
defrosted NTK, see neural tangent kernel
degradation problem, see also overly deep, see also residual network, 425, 426, 428, 431, 432, 435
degrees of freedom, 106, 403, 409
depth, 7, 40
determinant, 17
diagonalization, 17, 32, 118, 148
difference equation, see also training dynamics, 359, 360, 363–365
linear, 359, 362
  homogeneous, 359
inhomogeneous, 366
  nonlinear, 359
differential of the neural tangent kernel, see also meta kernel, 292–294, 339
collection to representation learning, 330
dNTK–preactivation cross correlation, see cross correlation
dynamical, see dynamical dNTK
iteration equation, see also forward equation, 297
name, 295
scaling laws, 311
dimensional analysis, 34, 34, 141, 311, 341, 401, 406
Dirac delta function, 50, 51, 76, 80, 118, 160, 171, 206, 213, 344, 392
integral representation, 50, 76
Dirac, Paul Adrien Maurice, ix, x, 191, 199
direct optimization, 321, 327, 357
directed acyclic graph, 41
discriminative model, see also probabilistic model, 192
disorder, see entropy
distillation, see knowledge distillation
dNTK, see differential of the neural tangent kernel
Don’t Panic, HHGTTG, 185
double factorial, 14
duality, 9, 242, 284, 286, 287, 289, 324, 328, 394
  learning algorithm – algorithm projectors, 383
  linear model – kernel methods, 282
  microscopic-macroscopic, 389, 394, 425
  nonlinear model – nearly-kernel methods, 317
dynamical dNTK, 319, 362, 369
dynamical NTK, see also effective kernel, see also interaction NTK, 317, 346, 359, 361
dynamics, see training dynamics
Eames (Inception meme), 165
diagonalization, 17, 32, 118, 148
difference equation, see also training dynamics, 359, 360, 363–365
linear, 359, 362
  homogeneous, 359
inhomogeneous, 366
  nonlinear, 359
differential of the neural tangent kernel, see also meta kernel, 292–294, 339
collection to representation learning, 330
dNTK–preactivation cross correlation, see cross correlation
dynamical, see dynamical dNTK
iteration equation, see also forward equation, 297
name, 295
scaling laws, 311
dimensional analysis, 34, 34, 141, 311, 341, 401, 406
Dirac delta function, 50, 51, 76, 80, 118, 160, 171, 206, 213, 344, 392
integral representation, 50, 76
Dirac, Paul Adrien Maurice, ix, x, 191, 199
direct optimization, 321, 327, 357
directed acyclic graph, 41
discriminative model, see also probabilistic model, 192
disorder, see entropy
distillation, see knowledge distillation
dNTK, see differential of the neural tangent kernel
Don’t Panic, HHGTTG, 185
double factorial, 14
duality, 9, 242, 284, 286, 287, 289, 324, 328, 394
  learning algorithm – algorithm projectors, 383
  linear model – kernel methods, 282
  microscopic-macroscopic, 389, 394, 425
  nonlinear model – nearly-kernel methods, 317
dynamical dNTK, 319, 362, 369
dynamical NTK, see also effective kernel, see also interaction NTK, 317, 346, 359, 361
dynamics, see training dynamics
Eames (Inception meme), 165
diagonalization, 17, 32, 118, 148
difference equation, see also training dynamics, 359, 360, 363–365
linear, 359, 362
  homogeneous, 359
inhomogeneous, 366
  nonlinear, 359
differential of the neural tangent kernel, see also meta kernel, 292–294, 339
collection to representation learning, 330
dNTK–preactivation cross correlation, see cross correlation
dynamical, see dynamical dNTK
iteration equation, see also forward equation, 297
name, 295
scaling laws, 311
dimensional analysis, 34, 34, 141, 311, 341, 401, 406
Dirac delta function, 50, 51, 76, 80, 118, 160, 171, 206, 213, 344, 392
integral representation, 50, 76
Dirac, Paul Adrien Maurice, ix, x, 191, 199
direct optimization, 321, 327, 357
directed acyclic graph, 41
discriminative model, see also probabilistic model, 192
disorder, see entropy
distillation, see knowledge distillation
dNTK, see differential of the neural tangent kernel
Don’t Panic, HHGTTG, 185
double factorial, 14
duality, 9, 242, 284, 286, 287, 289, 324, 328, 394
  learning algorithm – algorithm projectors, 383
  linear model – kernel methods, 282
  microscopic-macroscopic, 389, 394, 425
  nonlinear model – nearly-kernel methods, 317
dynamical dNTK, 319, 362, 369
dynamical NTK, see also effective kernel, see also interaction NTK, 317, 346, 359, 361
dynamics, see training dynamics
Eames (Inception meme), 165
diagonalization, 17, 32, 118, 148
difference equation, see also training dynamics, 359, 360, 363–365
linear, 359, 362
  homogeneous, 359
inhomogeneous, 366
  nonlinear, 359
Index

as an effective theory, 106
connection to RG flow, 106
in physics, 106
effective feature function, see feature function
effective kernel, see nearly-kernel methods
effective theory, 2, 43, 95, 105, 106, 125, 126,
138, 144, 145, 161, 164, 166, 203–205,
238, 240, 328, 373, 389, 395, 396,
407, 436
representation learning, 351
effective theory of deep learning, 43, 53, 64,
71, 73, 110, 164, 168, 192
effectively deep, see also optimal aspect ratio,
10, 336, 400, 423, 435
range extended by residual connections,
426
eigenvalue, 17
eightfold way, see also Gell-Mann, Murray, x
Einstein, Albert, 291
Einstein summation convention, 18
end of training, 262, 328, 335, 399
engineering, 383, 437
ensemble, see also probability distribution, 47,
123, 137, 145, 155, 165, 191, 192,
422, 435
entropy, 399, 400, 401–403, 405, 407, 409, 411,
414, 415, 418–420
additivity, 403, 404, 416
subadditivity, 403–405, 421
as a measure of disorder, 401
Boltzmann entropy, 402
Gibbs entropy, 402
next-to-leading-order correction, 417
Shannon entropy, 401
epigraph, 396
epilogue, 390, 396
epoch, see also stochastic gradient descent,
195
equivalence principle, 244, 245, 260, 272–275,
313, 315, 343, 358, 360, 425
connection to generalization, 273, 276
error factor, 197, 198, 201, 203, 241, 242, 250
ℓ-th-layer, 294, 337
cross-entropy loss, 250
MSE loss, 197
error function, 47
evidence, see Bayesian inference
expectation value, 11, 13, 21, 21, 227
exploding and vanishing gradient problem, 58,
122, 173, 227, 241, 242, 243, 244,
269, 425, 429, 431, 432, 436
connection to generalization, 274
for residual networks, 436
relation to criticality, 242
exploding and vanishing kernel problem, 58,
112, 113, 124, 171, 241–243, 245
expressivity, 40
extensivity of entropy, 403, 409
of loss, 162, 194, 267
extrapolation, see also *-polation, 277
Facebook AI Research, x, 177
FAIR, see Facebook AI Research
FCN, see fully-connected network
feature, see also representation, 42, 64, 105,
158, 186, 198, 200, 264, 353, 355, 389
vs. feature function, 289
feature function, 283, 288, 318, 333, 383
effective, 318, 323, 351
feature engineering for abstract inputs,
283
meta, see meta feature function
meta-meta, see meta-meta feature
function
nonlinear model, 318
random, see also random feature model,
8, 287–289, 392
feature indices, 318
feature space, 285, 325
feedforward network, 41, 182
ferromagnetism, 58
Feynman, Richard, 11
fine tuning, 59, 323
finite-width prediction
Bayesian inference, 184
finite-width prediction (cont.)

gradient descent, see also T-shirt equation, 373

fixed point

Banks–Zaks, see also optimal aspect ratio, 422

nontrivial, see also criticality, 58, 113, 125, 127, 128, 130, 131, 133, 141, 170

half-stable, see also GELU, see also SWISH, 135–137

trivial, 58, 112, 113, 123, 124, 136, 170, 172

float, see type (data)

fluctuations, 23, 54, 64, 110, 137, 154, 208, 381, 397, 425, 431

in deep linear networks, 63 vs. representation learning, 381

for your information, 422

force, see Newton’s second law

forward equation

\( \text{ddNTKs}, 385 \)
\( \text{dNTK}, 299 \)

general residual network, 436

MLP preactivations, 200, 202, 205–207, 299, 300

NTK, 200, 202, 203, 205, 207, 212, 215, 217, 243, 244

residual MLP preactivations, 428

forward pass, 205

four-point vertex, see also data-dependent coupling, 81, 100, 104, 137, 140, 142, 144, 145, 163, 182, 185, 209, 211, 220, 226, 227, 231, 232, 234, 237, 412, 421, 432

residual MLPs, 432

Fourier transform, 22

free energy, 412

frequentist probability, 155

frozen NTK, 228, 229, 231, 232, 234, 235, 238, 239, 243, 244, 249, 253, 258, 263, 267, 268, 270, 275, 276, 287, 311, 313, 350

\( \delta \) expansion, see \( \delta \) expansion

\( e_{1,2} \) expansion, see \( \delta \) expansion

features, 288

infinite-width limit of the NTK, 228

midpoint, 268, 272

polar angle parameterization, 274

full correlator, see correlator

fully-connected network, see also multilayer perceptron, 40

fully-trained condition

finite width, 348

infinite width, 253

function approximation, see also machine learning, 5, 38, 39, 40, 47, 53, 158, 160, 161, 168, 192, 193, 196–198

for linear models, 282

functional, 401, 404, 405

Gauss, Carl Friedrich, 247

Gauss–Jordan elimination, 177

Gaussian distribution, 8, 12, 21, 32, 34, 48, 54, 55, 66, 68, 137, 206, 211, 299, 392, 394, 396, 401, 413, 414

action, 27, 75

as a Gaussian process, 396

entropy, 409, 410

normal distribution, 13

relates to Dirac delta function, 50

single-variable, 13

zero-mean, defined by variance, 22

Gaussian expectation, see also bra-ket notation, 27, 79, 147–152, 185, 207, 208, 214–216, 220, 221, 229, 230, 234, 235, 237, 274, 413

Gaussian function, 12, 16, 39

Gaussian integral, 12

Gaussian process, see Gaussian distribution

gedanken inference, 168

gedanken model, 397

GELU, see activation function

general relativity, 17, 169, 213, 255

generalization, 166, 194, 196, 198, 264, 265, 411

generalization error, 264, 267, 272, 277, 390, 425

bias, 266–270, 379, 382

related to interpolation, 281

exact Bayesian inference, 270
Index

finite-width, 379, 383
optimal hyperparameter tuning, 273
robustness measure, 268
universality class analysis, 268
variance, 266, 267, 269, 271, 381
generalized posterior distribution, see posterior distribution
generating function, 14, 18, 19, 22, 212, 214, 215, 251, 302
giant leap, see also small step, 252, 348
Gibbs distribution, see Boltzmann distribution
Gibbs entropy, see entropy
Gibbs, J. Willard, 3
glasses (Bayesian), 169
goode odo calculation, 211
GPU, see graphical processing unit
gradient, 6
gradient clipping, see also exploding and vanishing gradient problem, 244
as Bayesian inference
at infinite width, 263
but not at finite width, 384
continuum or ODE limit, 258, 360, 372, 379, 380, 382
model fitting, 193, 195
stochastic, see stochastic gradient descent
tensorial, 196, 204, 254
wiring
finite width, 352, 353, 377
infinite width, 250
gradient-based learning, see gradient descent
graphical processing unit, 38
group representation theory, 105
Hamiltonian, see also neural tangent kernel, 192, 197, 360
hard drive, 401
hat (occupational), 401
Hebb, Donald, 180
Hebbian learning, see also neural association, 154, 179, 179, 182, 378, 400, 417
Herculean sequence, 148

Hessian, 6, 267
hidden layer, 41
hierarchical scaling, 34
Hinton, Geoffrey Everest, 227
Hopfield network, 182
Hubbard–Stratonovich transformation, 76, 213
human perception, 40
hybrid approach, 438
hype, 389
hyperparameters
architecture, see architecture hyperparameters
initialization, see initialization hyperparameters
regularization, see regularization hyperparameters
residual, see residual hyperparameters
scaling in an effective theory, 204
training, see training hyperparameters

Hyperparameters, see also hypothesis (Bayesian inference), 157
hypothesis (Bayesian inference), 154, 155–157, 167
categorical, see also cross-entropy loss, see also softmax distribution, 159, 160, 259, 260
deterministic, see also Dirac delta function, 158, 160
meta hypothesis, see meta hypothesis uncertain, see also Gaussian distribution, see also mean squared error, 158, 160, 260

identity matrix, 16
identity operator, 359
imaginary time, 360
indices
feature, see feature indices
layer, see layer indices
neural, see neural indices
sample, see sample indices
vectorial, see vectorial indices
induced distribution, 49, 51
inductive bias, 64, 111, 138, 154, 168, 168, 177–180, 182, 185, 242, 397, 400, 435 for representation learning in nonlinear models, 323 of activation functions, 277, 281 of learning algorithms, see also algorithm projector, 336, 372, 382, 383 of model architectures, 42, 333 of sparsity in deep learning, 397

Industrial Age, 2

infinite-width limit, see also not really deep, 7, 63, 64, 68, 72, 95, 137, 138, 166, 169, 177–179, 206, 209, 210, 221, 226, 228, 243, 249, 381, 396, 410, 421, 431 connection to linear models, 289 of deep linear networks, 62 of residual MLPs, 430

infinity, 247

InfoMax principle, see also unsupervised learning, 422

information, see also surprisal, 402, 403

Information Age, 3, 399

information theory, see also statistical mechanics, 10, 143, 382, 397, 399, 400, 402, 411 perspective on ∗-polation, 411 perspective on criticality, 410 perspective on generalization, 411

infrared (RG flow), 107

initialization (of you), 1, 399

initialization distribution, 5, 47, 55, 123, 137, 155, 157, 162, 165, 173, 182, 192, 193, 199, 201, 204, 205, 212, 391, 426

initialization hyperparameters, 48, 53, 59, 64, 110, 113, 124, 125, 127, 147, 153, 157, 162, 171, 227, 229, 234–236, 244, 258, 263, 264, 269, 273, 275, 292, 422, 425, 429, 437

critical, 58, 111, 114, 115, 123, 126–128, 131, 136–138, 141, 427 at finite width, 144, 145 for $K^* = 0$ universality, 131 for residual networks, 430

for scale-invariant universality, 123, 128 input data, see also test set, see also training set, see also validation set, 39, 42, 49, 55, 96, 97, 138, 157, 168–170, 183, 197, 201, 259, 283, 354, 394, 396, 417

int, see type (data)

integral representation, 51, 76

integrating out, see also marginalizing over, 80, 98, 162, 164, 165, 212–214, 221, 390

integration by parts, 122, 215, 229

intensity (of loss), 194

interacting theory, 9, see also non-Gaussian distribution, 32

entropy and mutual information, 411

variational method, 418

interaction NTK, 362, 363, 365, 366, 370

interactions, 8, 32, 33, 34, 63, 103, 110, 137, 179, 182, 186, 206, 207, 211, 396, 399, 404, 405

connection to statistical (in)dependence, 32

dynamics, 320, 360–362

weakly-interacting, 320

nearly-kernel methods, 328

self-interactions, 32

strong coupling, 396

interlayer correlation, 190, 211, 212, 214, 251, 252, 302

for dNTK evaluation, 301

interpolation, see also ∗-polation, 277

intralayer correlation, 211

inverse algorithm design, see also algorithm projector, 383

inverting tensor, 364, 365

iron, 58

irrelevant (RG flow), 108

Jacobian, 406

input-output, 138

Jaynes, Edwin T., 153, 155, 412

Jensen inequality, 404

Jenny B. Goode, 425

joules per kelvin (unit of entropy), 401
Index

k-nearest neighbors, see kernel methods
Kaiming initialization, see also initialization
hyperparameters, 125
kernel, see also metric, 100, 104, 111, 138,
227–229, 231, 232, 242, 244
δ expansion, see δ expansion
γ_0[a] basis, see γ_0[a] basis
effective kernel, see nearly-kernel methods
infinite-width limit of the metric, 100
kernel matrix
diagonal, 146, 150, 274
polar angle parameterization, 147, 274
linearized recursion, 112
meta kernel, see nearly-kernel methods
midpoint, 116, 117–120, 122, 123, 128,
130–133, 135, 136, 146
NTK, see neural tangent kernel
trained kernel, see nearly-kernel methods
kernel machine, see kernel methods
kernel methods, 286, 317, 325, 328
k-nearest neighbors, 286
as a memory-based method, 286
feature, see feature function
kernel, 285, 329, 383
Gaussian, 286
linear, 285
stochastic, 289
kernel trick, 286
prediction, 261, 287, 329, 375, 378
as a linear model, 282
stochastic kernel, see also random feature
model, 287
kernel trick, see kernel methods
kink, see also leaky ReLU, see also ReLU, 46
KL divergence, see Kullback-Leibler divergence
knowledge distillation, 177, 260
Konami Code, 168
Kronecker delta, 16, 48, 50, 57, 61, 66, 169,
196, 198, 209, 222, 402
Kullback-Leibler divergence, 259, 404, 405
kurtosis, excess, see also connected correlator,
24, 182
cokurtosis, 182
label, 192, 242, 260
hard, see also one-hot encoding, 260
soft, 260
label smoothing, see regularization
Landau, Lev, 3
language model, 42, 428, 436
Laplace transform, 22
Laplace’s principle of indifference, 402
large-n expansion, see 1/n expansion
layer, 4, 39
layer indices, 180, 181, 219, 279, 318, 409, 411
layer normalization, 437
lazy training, 355
leaky ReLU, see activation function
learning algorithm, see also Bayesian
inference, see also gradient descent,
5, 38, 54, 153, 161, 162, 168, 178,
191, 195, 196, 261, 262, 327, 383, 393
dual to algorithm projector, 383
learning rate, 195, 196, 198, 204, 234, 238, 240
global, 194, 196, 201, 203, 249, 253, 255, 393
step-dependent, 257
learning-rate tensor, 196, 200, 201, 204,
254, 255, 258, 293, 337, 353
layer-diagonal, 201
layer-independence, 294
learning rate equivalence principle, see
equivalence principle
Life, the Universe, & Everything, HHGTTG,
154
likelihood, see Bayesian inference
linear, see activation function
linear model, 264, 282, 283, 289, 317, 318,
355, 384, 392
for effective features, 319
is not a deep linear network, 289
linear regression, see also linear model, 283,
284, 318, 320
vs. quadratic regression, 320
linear transformations, 54, 55, 277
logistic function, see also softmax distribution,
44, 46, 47, 128, 159
loss, 160, 161, 191, 193, 194–197, 242
algorithm dependence at finite width, 373
loss (cont.)
auxiliary, 160
comparison of MSE and cross-entropy, 260
cross-entropy, 160, 242, 250, 258–261, 267, 404
MSE, 160, 193, 197, 203, 242, 250, 253, 254, 258, 265, 267, 353
for linear models, 283
generalized, 255, 352
name, 194
nonlinear models, 319
of generality, 265
SE, 194
test loss, 194, 197, 198, 266
relation to generalization, 265
training loss, 193, 194–198
relation to overfitting, 265
relation to underfitting, 265
lottery ticket hypothesis, 417, 423

machine learning, see also statistics (branch of mathematics), 39, 45, 155, 166, 191, 205, 261, 265, 282, 317, 390, 397
MacKay, David, 167, 389
macroscopic perspective, see also sample space, 2, 167, 390, 394, 396, 397, 399, 402, 425, 428
magic trick, 50, 212, 344
magnetic field, 58
magnetism, 58
MAP, see maximum a posteriori
marginal (RG flow), 108, 145, 350
marginalization rule, 96, 97, 98, 155
marginalizing over, see also integrating out, 80, 100, 164, 165, 207, 212
matrix-vector product, 177
matter, 3
maximum a posteriori, 161, 162, 165
gradient descent approximation, 262
maximum entropy, principle, 400, 402, 412, 421
maximum likelihood estimation, 161, 162, 165, 177, 194
gradient descent approximation, 262
Maxwell, James Clerk, 3
McFly, Martin Seamus “Marty”, 425
McGreevy, John, 71
mean, see also cumulant, see also moment, 13, 23
mean squared error, see loss
measurement precision cutoff, see cutoff
mechanics (physics), 191
memory-based method, see kernel methods, see nearly-kernel methods
meta feature function, 318, 333
dynamical, 336
random, 331
meta hypothesis, 166
meta kernel, see nearly-kernel methods
meta representation learning, see representation learning
meta-meta feature function, 319, 332
metric, see also data-dependent coupling, see also kernel, 74
first-layer, 74
infinite-width limit, 100
inverse, 75
ℓ-th-layer, 91
mean, 81
next-to-leading-order correction, 100, 103, 138, 140, 143–145, 350
second-layer, 81
stochastic, 81, 288
microscopic perspective, see also parameter space, 2, 383, 390, 394, 396, 399, 402, 425, 426
microstate (statistical mechanics), 412
midpoint input, 116, 116, 119, 122, 146, 268
midpoint kernel, see kernel
mini-batch, see batch
minimal model, 10
of deep learning, 43
of representation learning, see representation learning
Minsky, Marvin, 109, 227
MLE, see maximum likelihood estimation
MLP, see multilayer perceptron
mode, see also maximum a posteriori, 161
Index

model comparison
Bayesian, see Bayesian inference
linear model vs. quadratic model, 323
model complexity, 167, 322, 390, 391, 394–397
model fitting, see also training
Bayesian, see Bayesian inference
gradient-based optimization, see gradient descent
model parameters, see also biases, see also weights, 4, 38, 40, 49, 51, 165, 191–193, 195, 197, 389, 394, 425
connection to observables, 3
residual network, 436
molecule, 3
moment, see also full correlator, 14, 14, 18, 20, 21, 22, 24, 227
MSE, see mean squared error
MSE loss, see loss
multilayer perceptron, 40, 41, 76, 80, 157, 166, 168, 227, 241
a.k.a. a fully-connected network, 40
beyond, 436
vanilla, 427, 429
with residual connections, 427–429, 432, 436
mutual information, 399, 400, 405, 405, 407–411, 415, 417, 421, 422, 432
next-to-leading-order correction, 417
Narrator (Arrested Development), 166
nat (unit of entropy), 401
natural language processing, 42, 192, 389, 428, 436
natural logarithm, 401
naturalness, see also fine tuning, 323, 324
near-sparsity, see sparsity, principle of
action, 33, 88
as a nearly-Gaussian process, 396
connected correlators as observables, 26
entropy, 411
nearly-Gaussian process, see nearly-Gaussian distribution
nearly-kernel machine, see nearly-kernel methods
nearly-kernel methods, 292, 317, 327–329, 375
as a memory-based method, 328
effective kernel, 328–331
in terms of effective feature functions, 331
relation to dynamical NTK, 331
kernel, 325
meta kernel, 326
other potential names, 326
prediction, 327
trained kernel, see also trained NTK, 329, 375
prediction, 329
wiring, 329
nearly-linear model, see nonlinear model
nearly-linear regression, see quadratic regression
negative log probability, see action
negative log-likelihood, see also loss, 160, 161
neural association, see also Hebbian learning,
154, 179, 180, 182, 435
neural indices, 49, 57, 61, 66, 67, 76, 100, 137, 181, 197, 200, 202, 205–207, 215, 222, 249, 318
neural network, 1, 4, 37, 39, 42, 109, 191–193, 241, 389, 397
history, 38
neural tangent kernel, 139, 192, 197, 199, 204, 227, 228, 360
agitated, 228, 235, 239, 240
defined in conjunction with dNTK, 294
defrosted, 228
dynamical, see dynamical NTK
dynamics, 363
first-layer, 206, 207
frozen, see frozen NTK
interaction, see interaction NTK
ℓ-th-layer, 201, 202, 211, 212, 215, 219
mean, 211, 215, 216, 217, 220, 222–229, 234, 238, 239, 243
next-to-leading-order correction, 350
name, 197, 228, 360
NTK–preactivation cross correlation, see cross correlation
second-layer, 207, 208
step-independent, 359, 368
trained, see also trained kernel, 375
variance, 208–211, 220, 221, 222, 223, 226, 227, 231, 240
A-recursion, 224
B-recursion, 222
neuron, 1, 4, 39, 41
neuroscience, 37
Newton tensor, see also second-order update, 254, 255–257, 349
as a metric on sample space, 255, 352
generalized, 352
Newton’s method, 256, 256–259, 350, 382
as a second-order method, 256
Newton’s second law, 191
NLO metric, see metric
no-free-lunch theorem, 397
non-Abelian gauge theory, see also Banks–Zaks fixed point, 422
non-Gaussian distribution, see also nearly-Gaussian distribution, 31, 33, 68, 96
noninformative prior, see prior
nonparametric model, see also Gaussian process, 166, 396
nonlinear model, 292, 317, 318
cubic model, 319, 332
quadratic model, 292, 319, 322, 327, 330, 332
with wiring, 332
nontrivial fixed point, see fixed point
normal distribution, see Gaussian distribution
normalization factor, see also partition function, 13, 16, 27, 99, 118, 165, 166
not really deep, see also infinite-width limit, 10
NTK, see neural tangent kernel
objective function, see also loss, 193
observable, 3, 11, 14, 21, 155, 192, 197, 198, 245, 400, 403, 438
Occam’s razor, see also sparsity, principle of, 154, 166, 167, 168, 171, 323, 390, 402
ODE limit, see gradient descent
one-hot encoding, 170, 260
one-parameter families, 277, 431
optimal aspect ratio, see also effectively deep, 10, 336, 381, 400, 411, 421, 428, 432, 434, 435, 438
optimal brain damage, 417, 423
optimization, see gradient descent, see training, see also direct optimization, see also Newton’s method
orthogonal matrix, 16, 32
outcome space, 405, 407
output distribution, 49, 51, 64, 68, 158, 191
output matrix, 173
γ [ai] basis, see γ [ai] basis
overfitting, see also generalization, 166, 198, 265, 266, 390
by fine tuning the parameters, 323
overly deep, see also chaos, see also degradation problem, 10, 336, 400, 423, 425, 426
overparameterization, 166, 284, 285, 287, 389, 394, 397
in quadratic models, 321
Papert, Seymour, 109, 227
parallel susceptibility, 113, 121, 125, 171, 172, 229, 231, 233, 236, 237, 243, 269, 311, 430
paramagnetism, 58
parameter space, see also microscopic perspective, 195, 196, 254, 255, 336, 394
parameters, see model parameters
parity symmetry, 24, 25, 33
partition function, see also normalization factor, 15, 19, 27, 28, 29, 76, 92, 411
quadratic action, 27
with source, see also generating function, 14
perceptron, see Perceptron architecture
perceptron, see activation function
Perceptron architecture, 37, 38, 40
Index

permutation symmetry, 47, 123, 431
perpendicular susceptibility, 121, 122, 125, 171, 172, 229, 231, 233, 236, 244, 271, 311, 430
perturbation theory, 8, 11, 28, 31, 32, 118, 320, 360, 393, 411
perturbative cutoff, see cutoff
phase transition, 58
physics, 2, 3, 8, 71, 73, 76, 105, 125, 161, 181, 185, 320, 324, 344, 401
piece of cake, see also free dynamics, 359
point estimate, see also mode, 161
Polchinski, Joseph, 11
polynomial regression, 324
positive semidefinite matrix, 196
positive definite matrix, 16
posterior, 154, 156, 159–161, 163–166, 168, 169, 173, 176, 179, 182, 185, 191, 262, 263
generalized posterior distribution, see also gradient-based learning, 248, 262, 263, 266, 269, 411
infinite-width distribution, 176
posterior covariance, 175–177, 183, 185, 262, 263
finite width, 183
posterior mean, 176, 177, 183–186, 262, 384
finite width, 183
practical practitioners, 204, 238
preactivation, 39
pretraining, 11, 399, 422
principle, 2

criticality, see criticality
InfoMax, see InfoMax principle
learning-rate equivalence, see equivalence principle
maximum entropy, see maximum entropy, principle
of indifference, see Laplace’s principle of indifference
sparsity and near-sparsity, see sparsity, principle of
typicality, see typicality
variational, see variational principle

principles of deep learning theory, 43, 334
prior, 156, 157–159, 162, 163, 165–167, 169, 182, 191, 193, 400, 409
noninformation prior, see also Laplace’s principle of indifference, 402
probabilistic model, 155, 156, 159, 165, 166, 192
probability (branch of mathematics), see also Bayesian probability, see also frequentist probability, 11, 32, 155
probability distribution, 11, 12, 16, 18, 21, 22, 23, 26, 27, 47, 400
as a density, 406
programming, 39, 47
programming note, 173
PyTorch, 437
QED, 45
quadratic model, see nonlinear model
quadratic regression, see also quadratic model, 320, 320, 357
nearly-linear, 320
quantum electrodynamics, see QED
quantum mechanics, 3, 53, 118, 191, 199
Rabi, Isidor Isaac, 337
RAID, see also Redundant Array of Independent Disks, 403
random feature function, see feature function
random feature model, 287, 332
random meta feature model, 332
recurrent neural network, 241, 244
redundancy (information theory), 400, 408, 423
Redundant Array of Independent Disks, see also RAID, 403
regression, 260, 342
linear, see linear regression
nearly-linear, see quadratic regression
polynomial, see polynomial regression
regularization, 162, 260, 262, 323
early stopping, 260
for linear models, 284
interpretation of representation learning, 323
label smoothing, 260
regularization hyperparameters, 263
relative entropy, see Kullback-Leibler divergence
ReLU, see activation function
renormalization group flow, 105, 125, 144, 350, 389, 422
representation, see also feature, 73, 105, 137, 158, 179, 186, 200, 422
name, 105
of preactivations, 71
of the ddNTKs, 339
of the NTK, 296
of the NTK, 199
representation learning, 1, 8, 64, 169, 178, 179, 182, 185, 186, 188, 190, 261, 282, 289, 317, 334, 366, 381, 396, 422
as the evolution of feature functions, 289
for deep linear networks, 289
for quadratic models, 319, 322
manifested at finite width, 351
meta representation learning, 363
minimal model, 292, 317, 319, 329, 332, 333
nonminimal model, 319
vs. fluctuations, 381
vs. kernel learning, 290
residual block, 426, 427, 428, 435, 436
residual connection, 10, 43, 425, 426, 427, 429–432, 435
other names, 427
residual function, 426
residual hyperparameters, 428, 435, 436
optimal, 435
residual network, 43, 334, 381, 400, 423, 425, 426, 427, 429, 431, 436
general, 436
ResNet, 43, 428, 436
RG flow, see renormalization group flow, see representation group flow
RG flow and RG flow, 103, 126
RNN, see recurrent neural network
Rosenblatt, Frank, 37
Rumelhart, David Everett, 227
running coupling, 63, 64, 70, 98–100, 103, 105, 227, 415
quadratic, 98, 99, 415
quartic, 415
sextic, 418
saddle-point approximation, see also point estimate, 161
sample indices, 39, 49, 59, 65, 76, 98, 115, 137, 139, 159, 163, 169, 192, 197, 198, 206, 209, 210, 218, 229, 255
sample space, x, 254, 255, 349, 352, 394
saturation (of an activation), 45, 46, 243, 244
scale invariance, 45, 47, 113, 123, 125, 136, 137
scaling ansatz, 132, 134, 142, 143, 145, 151, 231, 239, 240, 311, 341
scaling hypothesis, 389, 390, 397
scaling law, 142, 231, 232, 240, 312, 316, 317, 380, 389, 417
Schrödinger’s cat, 155
Schwinger–Dyson equations, 86, 187, 329, 354
second-order method (optimization), see also Newton’s method, 254, 256
second-order update, see also Newton tensor, 254, 255, 352
generalized, 352, 353
self-averaging, see also Dirac delta function, 50, 82, 226, 249, 288
self-interaction, see interactions
semi-criticality, see criticality
semigroup, see also RG flow, 105
SGD, see stochastic gradient descent
Shannon entropy, see entropy
Shannon, Claude, 399
Shenker, Stephen, 71
shortcuts, see residual connection
sigmoid, see activation function
simple harmonic oscillator, see also Sho, x, 53

sin, see activation function

six-point vertex, see also data-dependent coupling, 395

skip connection, see residual connection

slay the beast (NTK variance), 221

small step, see also giant leap, 248, 252

softmax distribution, see also logistic function, 159, 160, 250, 259, 260

softplus, see activation function

source term, see also generating function, 14, 19, 212

spacetime, 153

sparsity, principle of, 8, 9, 166, 391, 397

near-sparsity at finite width, 392, 393, 396, 397

spin, see also bit (unit of entropy), 58, 115

spoiler alert, 190, 232

statement, see Bayesian probability

statistical dependence, see also interactions, see also nearly-Gaussian distribution, 34, 403, 417

statistical independence, 32, 32, 34, 63, 137, 177, 206, 208, 209, 403, 404, 409

absence of interactions and connection to Gaussian distribution, 32

statistical mechanics, see statistical physics

statistical physics, 3, 58, 110, 389, 400, 402, 412

statistics (branch of mathematics), see also machine learning, 161, 182, 390

statistics (of a random variable), see also probability distribution, 21

Bayesian interpretation, 155

steam engine, 2

step-evolution operator, 360, 361, 364, 368

stochastic gradient descent, 162, 195, 253, 257, 258

str, see type (data)

subleading corrections, see also 1/n expansion, 100, 101–103, 138, 143–145, 228

supervised learning, 192, 194, 197, 422

with linear models, see linear regression

with quadratic models, see quadratic regression

surprisal (information theory), 402, 405, 411

susceptibility

parallel, see parallel susceptibility

perpendicular, see perpendicular susceptibility

SWISH, see activation function

synergy (information theory), 408

T-shirt equation, 375

tablet, 230

tanh, see activation function

Taylor series, 5, 38

temperature, 58, 413

tensor, 5, 16, 28, 76, 196, 253

learning-rate tensor, see learning rate

Newton tensor, see Newton tensor

tensor decomposition

γ\^[a] basis, see γ\^[b] basis

ddNTKs R/S/T/U, 340, 377

dNTK-preactivation P/Q, 301, 306, 376

four-point correlator, 61

giving data-dependent couplings, 393

metric mean and fluctuation, 86, 187

NTK mean and fluctuation, 208, 215, 307, 308, 349

NTK variance A/B, 208, 222, 376

NTK–preactivation D/F, 210, 217

six-point correlator, 66

tensorial gradient descent, see gradient descent

test loss, see loss

test set, 194, 249, 256, 261, 264, 265, 396

thermodynamics, 2, 401, 412

traditionality, see also exploding and vanishing gradient problem, 244

trained kernel, see nearly-kernel methods

trained NTK, see neural tangent kernel

training, see also gradient descent, see also model fitting, 5, 39, 47, 162, 191, 193–195, 228, 241–244, 252

training data, see training set

training dynamics

controlled by the NTK, 192

finite width, 347–373
training dynamics (cont.)
inductive bias, 336
infinite width, 248, 250–256
training hyperparameters, 162, 195, 201,
202, 227, 228, 244, 258, 263,
264, 270, 272, 273, 275, 276,
293, 425, 429
independent from the optimization
algorithm, 358
training loss, see loss
training set, 5, 193, 194, 195, 198, 249,
253, 255, 261, 264, 390,
396
transformer, 42, 43, 157, 166, 168, 389, 428,
436, 437
transistor, 3
transition matrix, 90
translational invariance, 42, 55, 168
tripartite information, 408, 411, 423
trivial factor, see also exploding and
vanishing gradient problem,
241–243
trivial fixed point, see fixed point
truncated normal distribution, 48
Turing, Alan, 53
type (data)
floating-point precision, 407
integer, 34
string, 34
type safety, see also dimensional analysis, 34,
361
typicality, 63, 71, 137, 199
principle of, 165, 381
ultraviolet (RG flow), 107
underfitting, 265, 266
underparameterization, 284, 321
uniform distribution, 48, 413
universality, 106, 110, 125, 227, 389
of the fully-trained network solution, 348,
373, 382
universality class, 125
half-stable, 137

Index

$K^* = 0, 128, 130, 131, 134, 141–144, 152,$
227, 232, 233, 236, 238–240, 242, 243,
245, 270–272, 275, 312, 314, 342, 343,
421, 431, 433
scale-invariant, 126, 128, 132, 138–144,
146, 147, 150, 227, 232–236, 238,
242–244, 270, 274, 275, 312, 335, 344,
346, 421, 430, 433
transcended by scaling laws, 142
unstructured data, 397
unsupervised learning, 105, 382, 400, 411,
422, 428, 432, 434
as pretraining, 422
validation set, 265
variance, see also cumulant, 13, 16
variational ansatz, 412, 414
variational principle, see also maximum
entropy, principle, 400, 412,
412, 414
vectorial indices, 192, 198, 318
von Neumann, John, 1, 39
website, see deeplearningtheory.com
weight tying, see also convolutional neural
network, 42
weights, see also model parameters, 39, 193
Wick contraction, 20, 56, 57, 60, 61, 65, 74,
111, 185, 215, 222
Wick’s theorem, 11, 14, 15, 21, 28–30, 62, 65,
66
width, 7, 40
Williams, Ronald J., 227
wiring, see also Hebbian learning
in Bayesian inference, see Bayesian
inference
in gradient-based learning, see gradient
descent
in nearly-kernel methods, see
nearly-kernel methods
zero initialization, see also initialization
distribution, 47, 123, 431