References


© Springer Japan 2016
S. Amari, $\alpha$-divergence is unique, belonging to both $f$-divergence and Bregman divergence classes. IEEE Transactions on Information Theory, 55, 11, 4925–4931, 2009.


A. Cichocki and S. Amari, Families of $\alpha$, $\beta$- and $\gamma$-divergences: flexible and robust measures of similarities. Entropy, 12, 1532–1568, 2010.

A. Cichocki, S. Cruces and S. Amari, Generalized $\alpha$-$\beta$ divergences and their application to robust nonnegative matrix factorization. Entropy, 13, 134–170, 2011.

A. Cichocki, S. Cruces and S. Amari, Log-determinant divergences revisited: $\alpha$-$\beta$ and $\gamma$ log-det divergences. Entropy, 17, 2988–3034, 2015.


References

T. Ichimori, On rounding off quotas to the nearest integers in the problem of apportionment methods. JSIAM Letters, 3, 21–24, 2011.
References


Y. Ollivier, Riemannian metric for neural networks I: Feedforward networks. Information and
H. Park, S. Amari and K. Fukumizu, Adaptive natural gradient learning algorithms for various
M. Parry, A. P. Dawid and S. Lauritzen, Proper local scoring rule. Annals of Statistics, 40, 561–592,
2012.
G. Pistone, Examples of the application of nonparametric information geometry to statistical
G. Pistone and M. P. Rogantin, The exponential statistical manifold: mean parameters, orthogonality
G. Pistone and C. Sempo, An infinite-dimensional geometric structure on the space of all the prob-
C. R. Rao, Efficient estimates and optimum inference procedures in large samples. Journal of Royal
G. Raskutti and S. Mukherjee, The information geometry of mirror descent. IEEE Transactions on
J. Rauh, Finding the maximizers of the information divergence from an exponential family. IEEE
N. Ravishanker, E. L. Melnik and C. Tsai, Differential geometry of ARMA models. Journal of Time
A. Rényi, On measures of entropy and information, in Proc. 4th Symposium on Mathematical
N. L. Roux, P.-A. Manzagol and Y. Bengio, Topmoumoute online natural gradient algorithm. In
D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating
R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, Boosting the margin: A new explanation for
J. Schmidhuber, Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85–117,
2015.
J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University
S. Shinomoto, K. Shima and J. Tanji, Differences in spiking patterns among cortical neurons. Neural
P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, In
D. E. Rumelhart and J. L. McLelland (Eds.), Parallel Distributed Processing, 1, 194–281, MIT
A. Soriano and L. Vergara, Fusion of scores in a detection context based on alpha integration. Neural
T. Takenouchi and S. Eguchi, Robustifying AdaBoost by adding the naive error rate. Neural Com-
T. Takenouchi, S. Eguchi, N. Murata and T. Kanamori, Robust boosting algorithm against msla-


Index

Symbols
(α-β)-divergence, 100
(α, β)-divergence, 94
(α-β)-log-det divergence, 101
α-divergence, 58, 67, 72
α-expert machine, 84
α-family of probability distributions, 81
α-function, 57
α-geodesic, 75
α-geometry, 136
α-integration, 82
α-mean, 77
α-projection theorem, 76
α-Pythagorean theorem, 76
β-divergence, 95
χ-divergence, 91
χ-escort distribution, 91
χ-exponential family, 90
e-affine parameter, 38
e-condition, 257
e-flat, 38
e-geodesic, 38
e-parallelogram transport, 202
em algorithm, 28
(F, G, H)-structure, 104
f-divergence, 54
γ-divergence, 102
k-cut, 144
k-means, 234
k-sparse, 336
κ-exponential family, 89
L0-norm, 338
L1-norm, 338
m-affine parameter, 38
m-condition, 257
m-flat, 38
m-geodesic, 38
m-parallel transport, 202
m-projection, 46, 252
φ-center of cluster, 233
Φ-function method, 245
q-divergence, 85
q-entropy, 85
q-escort geometry, 92
q-exponential, 85
q-exponential family, 86, 89
q-free energy, 87
q-logarithm, 85
q-metric, 88
(ρ, τ)-structure, 104
U-divergence, 95
(u, v)-divergence, 92
(u, v)-structure, 99

A
Absolute-value-based Hessian natural gradient, 286
Active set, 340
Adaptive learning method, 294
Adaptive natural gradient learning, 293
Affine connection, 112
Affine coordinate system, 18
Affine flat structure, 19
Akaike information criterion, 312
Alternating minimization algorithm, 27
Amari–Chentsov structure, 134
Amari–Chentsov tensor, 134
Ancillary submanifold, 169
Ancillary tangent subspace, 201
ARMA model, 218
AR model, 217
Asymptotic theory of hypothesis testing, 175
Auto-correlation coefficients, 221

© Springer Japan 2016
S. Amari, Information Geometry and Its Applications,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8
Auto-regression model, 217

B
Back-propagation learning, 281
Barrier function, 345
Basis vectors, 20
Bayesian duality, 266
Bayesian posterior distribution, 266
Belief propagation, 249
Blow-down technique, 309
Boltzmann machine, 181, 268
Boosting, 261
Bregman divergence, 13

C
Canonical divergence, 138
Canonical parameter, 32
Central limit theorem, 60
Chernoff divergence, 242
Chernoff information, 242
Clique, 250
Clustering, 231
Clustering algorithm, 234
Coarse graining, 53
Cocktail party problem, 323
Coefficient of proportionality, 191
Coefficients of affine connection, 113
Conformal transformation, 91
Conformal transformation of a kernel, 249
Conjugate priors, 267
Consistent estimator, 168
Contrastive divergence, 273
Convex-concave computational procedure, 249
Convex function, 12
Convex programming, 345
Coordinate system, 4
Coordinate transformation, 4
Covariant derivative, 117
Cramèr–Rao bound, 166
Cramèr–Rao theorem, 166
Critical region, 300
Critical slowdown, 305
Cubic tensor, 115
Cumulant generating function, 32

D
Decomposable divergence, 55
Deep learning, 292, 296
Deformed exponential family, 89
Divergence, 9

Dual affine structure, 19
Dual connections, 131
Dual convex function, 17
Dual geodesic, 19
Dually flat manifold, 137

E
Efficient, 173
Efficient score, 194
Einstein summation convention, 20
Eliminating singularity, 299
EM algorithm, 179
Embedding curvature, 129, 349
Ergodic time series, 215
Error covariance matrix, 166
Escort probability distribution, 88
Estimating function, 197
Estimator, 165
Euler–Schouten curvature, 129
Exponential family, 31

F
First-order asymptotic theory, 173
Fisher information matrix, 33
Foliation, 145
Free energy, 32

G
Game, 349
Game-divergence, 350
Game-score, 349
Gaussian kernel, 247
Gaussian mixture model, 180
Gaussian RBM, 275
Generalization error, 280
Generalized inverse, 337
Generalized Pythagorean theorem, 24
Geodesic, 19, 117
Graph Laplacian, 356
Graphical model, 250

H
Hidden variable, 179
Higher-order asymptotic theory of estimation, 173
Higher-order correlations, 149
Higher-order cumulants, 326
Hyvärinen divergence, 354
Hyvärinen score, 353
<table>
<thead>
<tr>
<th>Index</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Independent component analysis, 322</td>
<td></td>
</tr>
<tr>
<td>Information integration, 150</td>
<td></td>
</tr>
<tr>
<td>Information monotonicity, 52</td>
<td></td>
</tr>
<tr>
<td>Inner product, 23</td>
<td></td>
</tr>
<tr>
<td>Input–output analysis, 157</td>
<td></td>
</tr>
<tr>
<td>Instantaneous loss, 280</td>
<td></td>
</tr>
<tr>
<td>Integrated information, 152</td>
<td></td>
</tr>
<tr>
<td>Integration of weak machines, 261</td>
<td></td>
</tr>
<tr>
<td>Invariance criterion, 51</td>
<td></td>
</tr>
<tr>
<td>Invariant divergences, 52</td>
<td></td>
</tr>
<tr>
<td>Invariant Riemannian metrics, 52</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
</tr>
<tr>
<td>Kernel exponential family, 42</td>
<td></td>
</tr>
<tr>
<td>Kernel function, 246</td>
<td></td>
</tr>
<tr>
<td>Killing metric, 329</td>
<td></td>
</tr>
<tr>
<td>KL-divergence, 71, 220</td>
<td></td>
</tr>
<tr>
<td>Kronecker-factored approximate curvature, 293</td>
<td></td>
</tr>
<tr>
<td>Kullback–Leibler (KL) divergence, 11</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Large deviation, 60</td>
<td></td>
</tr>
<tr>
<td>Large deviation theorem, 61</td>
<td></td>
</tr>
<tr>
<td>Learning constant, 294</td>
<td></td>
</tr>
<tr>
<td>Least angle regressions, 343</td>
<td></td>
</tr>
<tr>
<td>Least equiangle theorem, 342</td>
<td></td>
</tr>
<tr>
<td>Legendre transformation, 16</td>
<td></td>
</tr>
<tr>
<td>Levi–Civita transformation, 16</td>
<td></td>
</tr>
<tr>
<td>Linear machine, 242</td>
<td></td>
</tr>
<tr>
<td>Linear system, 215</td>
<td></td>
</tr>
<tr>
<td>Loss of information by data reduction, 185</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Machine learning, 231</td>
<td></td>
</tr>
<tr>
<td>MA model, 218</td>
<td></td>
</tr>
<tr>
<td>Manifold, 3</td>
<td></td>
</tr>
<tr>
<td>Margin, 243</td>
<td></td>
</tr>
<tr>
<td>Maximum entropy, 223</td>
<td></td>
</tr>
<tr>
<td>Maximum entropy principle, 45</td>
<td></td>
</tr>
<tr>
<td>Maximum likelihood estimator, 48</td>
<td></td>
</tr>
<tr>
<td>Mean field approximation, 254</td>
<td></td>
</tr>
<tr>
<td>Metric affine connection, 125</td>
<td></td>
</tr>
<tr>
<td>Milnor attractor, 310</td>
<td></td>
</tr>
<tr>
<td>Minimum description length, 312</td>
<td></td>
</tr>
<tr>
<td>Minimum entropy, 224</td>
<td></td>
</tr>
<tr>
<td>Minkovskian gradient, 343</td>
<td></td>
</tr>
<tr>
<td>Minor subspace, 317</td>
<td></td>
</tr>
<tr>
<td>Mirror descent method, 289</td>
<td></td>
</tr>
<tr>
<td>Misspecified model, 186</td>
<td></td>
</tr>
<tr>
<td>Mixed coordinate system, 144</td>
<td></td>
</tr>
<tr>
<td>Mixture family, 37</td>
<td></td>
</tr>
<tr>
<td>Moving-average model, 218</td>
<td></td>
</tr>
<tr>
<td>Multilayer perceptron, 292, 296</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Natural gradient, 283</td>
<td></td>
</tr>
<tr>
<td>Natural gradient learning method, 284</td>
<td></td>
</tr>
<tr>
<td>Natural parameter, 32</td>
<td></td>
</tr>
<tr>
<td>Natural policy gradient, 288</td>
<td></td>
</tr>
<tr>
<td>Negative entropy, 33</td>
<td></td>
</tr>
<tr>
<td>Neyman–Scott problem, 191</td>
<td></td>
</tr>
<tr>
<td>Non-holonomic coordinate system, 329</td>
<td></td>
</tr>
<tr>
<td>Non-negative matrix factorization, 333</td>
<td></td>
</tr>
<tr>
<td>Nuisance parameter, 191</td>
<td></td>
</tr>
<tr>
<td>Nuisance tangent subspace, 201</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Observed point, 47</td>
<td></td>
</tr>
<tr>
<td>Observed submanifold, 180</td>
<td></td>
</tr>
<tr>
<td>On-line learning, 281</td>
<td></td>
</tr>
<tr>
<td>Overlapping singularity, 299</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Parallel transport, 22, 118</td>
<td></td>
</tr>
<tr>
<td>Parameter of interest, 191</td>
<td></td>
</tr>
<tr>
<td>Plateau, 302</td>
<td></td>
</tr>
<tr>
<td>Plateau phenomena, 308</td>
<td></td>
</tr>
<tr>
<td>Policy natural gradient, 284</td>
<td></td>
</tr>
<tr>
<td>Polynomial kernel, 247</td>
<td></td>
</tr>
<tr>
<td>Positive-definite symmetric matrix, 96</td>
<td></td>
</tr>
<tr>
<td>Power spectrum, 217</td>
<td></td>
</tr>
<tr>
<td>Principal component, 317</td>
<td></td>
</tr>
<tr>
<td>Principal component analysis, 315</td>
<td></td>
</tr>
<tr>
<td>Principal subspace, 317</td>
<td></td>
</tr>
<tr>
<td>Prior distribution, 266</td>
<td></td>
</tr>
<tr>
<td>Projection theorem, 25, 143</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
</tr>
<tr>
<td>RAS transformation, 160</td>
<td></td>
</tr>
<tr>
<td>RC curvature, 119</td>
<td></td>
</tr>
<tr>
<td>Reinforcement learning, 287</td>
<td></td>
</tr>
<tr>
<td>Restricted Boltzmann machine, 268</td>
<td></td>
</tr>
<tr>
<td>Riemann–Christoffel curvature tensor, 119</td>
<td></td>
</tr>
<tr>
<td>Riemannian connection, 113</td>
<td></td>
</tr>
<tr>
<td>Riemannian geometry, 109</td>
<td></td>
</tr>
<tr>
<td>Riemannian gradient, 283</td>
<td></td>
</tr>
<tr>
<td>Riemannian metric, 19</td>
<td></td>
</tr>
<tr>
<td>Riemannian structure, 10</td>
<td></td>
</tr>
<tr>
<td>Robust cluster center, 238</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
</tr>
<tr>
<td>Saddle-free Newton method, 286</td>
<td></td>
</tr>
<tr>
<td>Scale problem, 191</td>
<td></td>
</tr>
<tr>
<td>Score function, 111</td>
<td></td>
</tr>
<tr>
<td>Semi-definite programming, 346</td>
<td></td>
</tr>
<tr>
<td>Shape parameter, 212</td>
<td></td>
</tr>
<tr>
<td>Singular point, 301</td>
<td></td>
</tr>
<tr>
<td>Singular prior, 313</td>
<td></td>
</tr>
<tr>
<td>Singular statistical models, 311</td>
<td></td>
</tr>
<tr>
<td>Singular structure, 296</td>
<td></td>
</tr>
<tr>
<td>Soft clustering, 236</td>
<td></td>
</tr>
<tr>
<td>Solution path, 341</td>
<td></td>
</tr>
<tr>
<td>Sparse vector, 336</td>
<td></td>
</tr>
<tr>
<td>Standard estimating function, 332</td>
<td></td>
</tr>
<tr>
<td>Standard ( f )-divergence, 56</td>
<td></td>
</tr>
<tr>
<td>Stiefel manifold, 320</td>
<td></td>
</tr>
<tr>
<td>Stochastic descent learning method, 281</td>
<td></td>
</tr>
<tr>
<td>Stochastic relaxation, 286</td>
<td></td>
</tr>
<tr>
<td>Submanifold, 126</td>
<td></td>
</tr>
<tr>
<td>Sufficient statistic, 52</td>
<td></td>
</tr>
<tr>
<td>Super efficiency, 332</td>
<td></td>
</tr>
<tr>
<td>Support vector, 244</td>
<td></td>
</tr>
<tr>
<td>Support vector machine, 242</td>
<td></td>
</tr>
<tr>
<td>System complexity, 152</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tangent space, 19, 109</td>
</tr>
<tr>
<td>Tangent subspace of interest, 201</td>
</tr>
<tr>
<td>Temporal firing pattern, 211</td>
</tr>
<tr>
<td>Tensor, 114</td>
</tr>
<tr>
<td>Time series, 215</td>
</tr>
<tr>
<td>Total Bregman divergence, 238</td>
</tr>
<tr>
<td>Total least squares, 196</td>
</tr>
<tr>
<td>Training error, 280</td>
</tr>
<tr>
<td>Transfer function, 217</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidentifiability, 298</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voronoi diagram, 234</td>
</tr>
</tbody>
</table>