The calculus of deterministic equivalents and its applications to high-dimensional statistics

Edgar Dobriban

University of Pennsylvania

September 15, 2021

Overview

Outline

Motivation

- Standard linear model $Y = X\beta + \varepsilon$, where
  1. Y is $n \times 1$ outcome, X is $n \times p$ feature matrix.
  2. β is p-dim parameter
- Ordinary least squares

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

- Mean squared error of OLS, assuming $\mathbb{E}\varepsilon = 0$, $\operatorname{cov}(\varepsilon) = \sigma^2 I_n$

$$\mathbb{E}\|\hat{\beta} - \beta\|^2 = \sigma^2 \operatorname{tr}[(X^\top X)^{-1}]$$

- How large is this?

Motivation ctd

- When $X_{ij} \sim \mathcal{N}(0,1)$ are iid standard normal,

$$\mathbb{E}\,\mathrm{tr}[(X^\top X)^{-1}] = \frac{p}{n-p-1}.$$

- More general data distributions? There are only approximate expressions.
- $x_i = \Sigma^{1/2} z_i \in \mathbb{R}^p$, where z_i have independent standardized entries, for $i = 1, \ldots, n$. Then with $\widehat{\Sigma} = n^{-1} X^\top X$,

$$\widehat{\Sigma}^{-1} \asymp \frac{n}{n-p} \cdot \Sigma^{-1}.$$

  and $\mathrm{tr}[(X^\top X)^{-1}] \asymp \frac{p}{n-p} \cdot \mathrm{tr}\,\Sigma^{-1}/p$.

Calculus of deterministic equivalents

- Deterministic equivalents are a powerful tool in random matrix theory (Serdobolskii 1980s, Hachem et al 2007, etc). Here we develop a systematic approach.
- We have sequences of (not necessarily symmetric) $k_n \times k_n$ random matrices A_n and deterministic matrices B_n of growing dimensions
- **Definition**: B_n is a *deterministic equivalent* for A_n,

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} |\mathrm{tr}(C_n A_n) - \mathrm{tr}(C_n B_n)| = 0$$

almost surely, for any $k_n \times k_n$ sequence C_n of (not necessarily symmetric) deterministic real matrices with bounded trace norm, i.e.,

$$\limsup_{n\to\infty} \|C_n\|_{tr} = \limsup_{n\to\infty} \sum_i |\sigma_i(C_n^\top C_n)^{1/2}| < \infty.$$

e.g, $C_n = c_n c_n^\top$, $\|c_n\|_2$ bounded

Calculus of deterministic equivalents

- $\mathrm{tr}(C_n A_n)$ is a linear combination of entries of A_n
- $A_n \asymp B_n$ if each linear combination of entries of A_n can be approximated by the same linear combination of entries of B_n

Sample covariance matrices

Example 1. (Mestre et al., 2011)

Let $\hat{\Sigma} = X^\top X/n$, where $X = Z\Sigma^{1/2}$ and Z is an $n \times p$ random matrix with iid entries of zero mean, unit variance and finite $8 + \eta$ moment. Also, $\Sigma^{1/2}$ is any sequence of $p \times p$ positive semi-definite matrices satisfying $\sup \|\Sigma\|_2 < \infty$. As $n, p \to \infty$ proportionally, for any $\lambda > 0$

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (q_p \Sigma + \lambda I_p)^{-1},$$

where q_p is the solution of a fixed point equation.

1. Similar results for elliptical model, where datapoints can have different scalings: $x_i = g_i^{1/2}\Sigma^{1/2}z_i$.
2. This is the simplest way I know how to think of a broad class of results in random matrix theory.

Rules of calculus

The calculus of deterministic equivalents has the following properties.

1. **Sum.** If $A_n \asymp B_n$ and $C_n \asymp D_n$, then $A_n + C_n \asymp B_n + D_n$.
2. **Product.** If $\limsup \|A_n\|_{op} < \infty$, A_n is independent of B_n, C_n, and $B_n \asymp C_n$, then $A_n B_n \asymp A_n C_n$.
3. **Trace.** If $A_n \asymp B_n$, then $\mathrm{tr}\{k_n^{-1} A_n\} - \mathrm{tr}\{k_n^{-1} B_n\} \to 0$ almost surely.
4. **Derivative.** If $f(A_n, z) \asymp g(B_n, z)$, for analytic f, g on an open domain of $\mathbb{C}$, then $\partial_z f(A_n, z) \asymp \partial_z g(B_n, z)$.

Outline

Setup

- Standard linear model $Y = X\beta + \varepsilon$
- Samples distributed across k machines. The i-th machine has matrix X_i ($n_i \times p$) and outcomes Y_i.

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_k \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_k \end{bmatrix}$$

- Global least squares - infeasible
- *Local* least squares estimator $\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top Y_i$ (assume $n_i > p$)
- Send to parameter server, average
- How does this compare to OLS on full data?

Parameter estimation

- Weighted distributed estimator, $\sum_{i=1}^k w_i = 1$

$$\hat\beta_{dist} = \sum_{i=1}^k w_i \hat\beta_i.$$

- MSE on i-th machine is

$$\mathbb{E}\|\hat\beta_i - \beta\|^2 = \sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]$$

- Optimal "inverse variance weighting": $w_i^* \propto 1/[\sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]]$
- *Relative efficiency*

$$RE(X_1, \ldots, X_k) = \frac{\mathbb{E}\|\hat\beta - \beta\|^2}{\mathbb{E}\|\hat\beta_{dist} - \beta\|^2} = \operatorname{tr}[(X^\top X)^{-1}] \left[\sum_{i=1}^k \frac{1}{\operatorname{tr}[(X_i^\top X_i)^{-1}]} \right]$$

How does this depend on n, p, k?

Discoveries under asymptotics

- CDE: $\operatorname{tr}[(X_i^\top X_i)^{-1}] \asymp \frac{p}{n_i - p} \cdot \operatorname{tr} \Sigma^{-1}/p$.
  The RE has a simple approximation (n samples, p dimensions, k machines)
  $$\frac{\mathbb{E}\|\hat{\beta} - \beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2} \approx \frac{n - kp}{n - p}$$
- Can be computed conveniently in practice. e.g., $n = 10^9$, $p = 10^6$, $k = 100$, then $RE \approx 10/11 \approx 0.91$, so we keep 90% efficiency

A general framework

- Important to study not only estimation, but also prediction/test error, residual error, confidence intervals etc
- Predict the linear functional

$$L_A = A\beta + Z$$

- Using the plug-in estimator

$$\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$$

- A - fixed $d \times p$ matrix; mean and covariance of Z has the structure: $Z \sim (0, h\sigma^2 I_d)$, $h \geqslant 0$
- The noise can be correlated with ε: $\mathrm{Cov}\left[\varepsilon, Z\right] = N$ (e.g., to study residuals)
- Relative efficiency:

$$E(A; X_1, \ldots, X_k) := \frac{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta})\|^2}{\mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_{dist})\|^2}.$$

Examples: Predict $L_A = A\beta + Z$ by $\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0$

Statistical learning problem	L_A	$\hat{L}_A$	A	h	N
Parameter estimation	β	$\hat{\beta}$	I_p	0	0
Regression function estimation	$X\beta$	$X\hat{\beta}$	X	0	0
Confidence interval for marginal effect	β_j	$\hat{\beta}_j$	$E_j^\top$	0	0
Test error	$x_t^\top \beta + \varepsilon_t$	$x_t^\top \hat{\beta}$	$x_t^\top$	1	0
Training error/Residual	$X\beta + \varepsilon$	$X\hat{\beta}$	X	1	$\sigma^2 I_n$

Finite sample results

- When $h = 0$ (no noise), the MSE of estimating $L_A = A\beta$ by OLS $\hat{L}_A = A\hat{\beta} = A(X^\top X)^{-1}X^\top Y$ is

$$M(\hat{\beta}) = \sigma^2 \cdot \text{tr}\left[(X^\top X)^{-1}A^\top A\right].$$

- For the distributed estimator $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, $\sum_i w_i = 1$

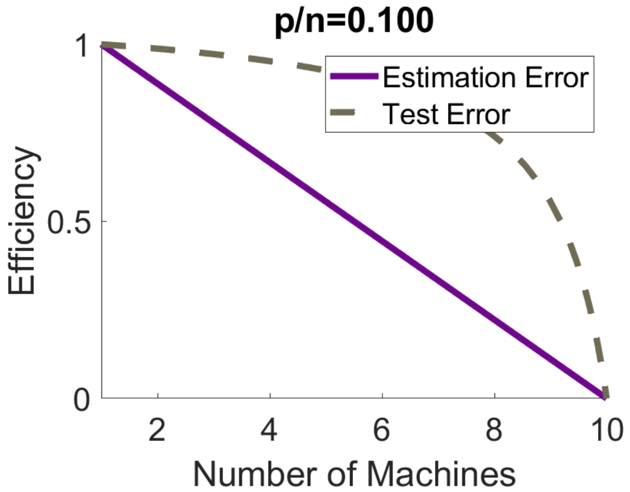
$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \sum_{i=1}^{k} w_i^2 \cdot \text{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right].$$

- So optimal efficiency is

$$E(A; X_1, \ldots, X_k) = \text{tr}\left[(X^\top X)^{-1}A^\top A\right] \cdot \sum_{i=1}^{k} \frac{1}{\text{tr}\left[(X_i^\top X_i)^{-1}A^\top A\right]}.$$

CDE: $\text{tr}[(X_i^\top X_i)^{-1}A^\top A] \asymp \frac{p}{n_i - p} \cdot \text{tr}[\Sigma^{-1}A^\top A]/p$.

Plot efficiencies



The loss of efficiency is much worse for estimation ($\frac{\mathbb{E}\|\hat{\beta}-\beta\|^2}{\mathbb{E}\|\hat{\beta}_{dist}-\beta\|^2}$) than for test error ($\frac{\mathbb{E}(x_t^\top\hat{\beta}-y_t)^2}{\mathbb{E}(x_t^\top\hat{\beta}_{dist}-y_t)^2}$).

Outline

Distributed ridge regression

- Global ridge estimator $\hat{\beta}(\lambda) = (X^\top X + n\lambda I_p)^{-1} X^\top Y$
- Local ridge estimator $\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top Y_i$
- One-shot weighted estimator

$$\hat{\beta}_{dist}(w) = \sum_{i=1}^{k} w_i \hat{\beta}_i$$

- Key point: no constraints on the weights w because ridge estimator is biased!
- This leads to some surprising consequences, e.g. optimal weights do not sum to unity
- Also, do not require $n_i > p$ anymore

Optimal weights and MSE

- Goal: find optimal weights w to minimize $\mathbb{E}\|\hat{\beta}_{dist}(w) - \beta\|^2$
- Optimal weights $w^* = (A + R)^{-1} v$, where $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i$

$$v_i = \beta^\top Q_i \beta, \quad A_{ij} = \beta^\top Q_i Q_j \beta, \quad R_{ii} = \frac{\sigma^2}{n_i} \mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]$$

- Assume β is random and independent of ε. Mean and variance: $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma^2$, $\mathbb{E}\beta_i = 0$, $\mathbb{E}\beta_i^2 = \sigma^2\alpha^2/p$
- Concentration of quadratic forms:

$$\beta^\top M \beta - \frac{\alpha^2\sigma^2}{p} \cdot \mathrm{tr}(M) \to_{a.s.} 0$$

- Need to know the limits of

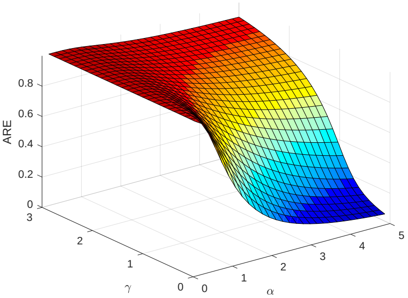
$$\mathrm{tr}\, Q_i = \mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i], \quad \mathrm{tr}\, Q_i Q_j, \quad R_{ii}$$

Use product rule and differentiation rule of CDE

Findings

- How to choose the tuning parameters λ_i? When $\Sigma = I$, the MSE decouples over k machines, which means locally optimal λ_i are also globally optimal!
- Again, when $\Sigma = I$, the efficiency is positive when $k \to \infty$ – *infinite-worker limit*

Landscape of RE for infinite-worker limit



- Suggests that one-shot learning is practical and has good performance
- In the "low dimension and high SNR" region, one should use other methods, e.g. iterative ones

Outline

Bias-variance decomposition

Choose $\hat{f}$ based on the training set, and decompose the test error into bias and variance ($\mathbb{E}_{x,y} = \mathbb{E}_{(x,y)\sim\text{test}}$):

$$\begin{aligned}\mathbb{E}_{x,y}\mathbb{E}(y - \hat{f}(x))^2 &= \mathbb{E}_{x,y}\mathbb{E}(y - \mathbb{E}\hat{f}(x))^2 + \mathbb{E}_{x,y}\mathrm{Var}(\hat{f}(x)) \\ &= \mathrm{Bias}^2 + \mathrm{Variance}.\end{aligned}$$

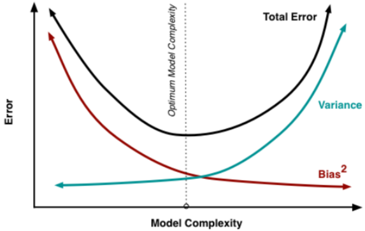
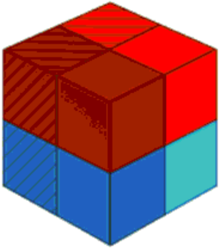


Figure: Bias and variance contributing to total error.[1]

[1]Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Our approach

- Variance depends on randomness in: initialization, input features, outcomes/labels...
  and other aspects: randomness in optimization algorithm, ...
- Decompose the variance into its ANOVA components (R.A. Fisher, 1918)



Three-way ANOVA: how is a response affected by three factors?[2]

[2]Figure: www.graphpad.com/guides/prism/latest/statistics

Setup

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where x has i.i.d. standardized entries, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is label noise:

$$Y = X\theta + \mathcal{E}.$$

- **Training**: Fit a two-layer linear (later nonlinear) random features model

$$f(x) = (Wx)^\top \beta.$$

- Weights $W \in \mathbb{R}^{p \times d}$, $p \leq d$ drawn uniformly from partial orthonormal matrices, $WW^\top = I_p$. Assume $\theta \sim \mathcal{N}(0, \alpha^2 I_d/d)$. Train β with L_2 loss, L_2 regularization λ to get predictor:

$$f(x) = (Wx)^\top \hat{\beta}_{\lambda, \mathcal{T}, W} = x^\top W^\top \left(\frac{WX^\top XW^\top}{n} + \lambda I_p \right)^{-1} \frac{WX^\top Y}{n}.$$

ANOVA: Symmetric variance decomposition

Denote $(X, W, \mathcal{E})$ by (s, i, l) respectively. We decompose the variance of $\hat{f}$ in a symmetric way via the analysis of variance (ANOVA):

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}
V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s, l, i\} \\
V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a, b)] - V_a - V_b, && a, b \in \{s, l, i\}, a \neq b. \\
V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a, b, c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} \\
&= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a, b, c\} = \{s, l, i\}.
\end{aligned}$$

- V_a: the effect of varying a alone *(main effect)*.
- V_{ab}: the second-order *interaction effect* between a and b beyond their main effects.
- V_{abc}: interaction effect among a, b, c beyond their pairwise interactions.

Calculation of the variance components

Define

$$\tilde{M} := W^\top (n^{-1} W X^\top X W^\top + \lambda I_p)^{-1} W X^\top / n \qquad M := \tilde{M} X.$$

Then we have $\hat{f}(x) = x^\top \tilde{M} Y = x^\top M\theta + x^\top \tilde{M}\mathcal{E}$. For V_s,

$$\begin{aligned} V_s &= \mathbb{E}_{\theta,x} \mathrm{Var}_X(\mathbb{E}_{\mathcal{E},W}(\hat{f}(x)|X)) = \mathbb{E}_{\theta,x,X}[x^\top(\mathbb{E}_W M - \mathbb{E}M)\theta]^2 \\ &= \frac{\alpha^2}{d} \mathbb{E}_X \|\mathbb{E}_W M - \mathbb{E}M\|_F^2. \end{aligned}$$

Evaluate $\mathbb{E}M$ in two steps:

1. When X is Gaussian, express in terms of eigenvalues of $\widehat{\Sigma}$; then use Marchenko-Pastur theorem.
2. Let $\tilde{R}_\tau = \left(\frac{X(W^\top W + \tau)X^\top}{n} + \lambda I_n\right)^{-1}$, $M_\tau = W^\top W X^\top \tilde{R}_\tau \frac{X}{n}$.

(1). *CDE* : $\quad \lim_{\tau\to 0} \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\, M_\tau M_\tau^\top / d \quad$ is independent of the dist. of X.

(2). $\quad \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(MM^\top)/d = \lim_{\tau\to 0} \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\, M_\tau M_\tau^\top / d$

Calculus of deterministic equivalents for Haar projections

Example 2. (Couillet et al., 2012)

Let $W \in \mathbb{R}^{p\times d}$ be the first p rows of a unitary Haar distributed random matrix. Suppose $R^{d\times d}$ is a sequence of positive semi-definite random matrices such that $\sup \|R\|_2 < \infty$, almost surely. As $p, d \to \infty$ proportionally, for any $\lambda > 0$

$$(R^{1/2} W^\top W R^{1/2} + \lambda I_d)^{-1} \overset{w}{\asymp} (\bar{e}_d R + \lambda I_d)^{-1},$$

where $\bar{e}_d$ is the solution of a fixed point equation.

Main result: ANOVA for two-layer linear NN

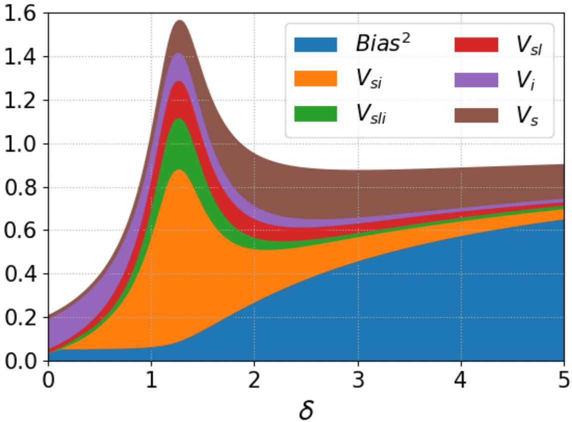
- Asymptotics: $d \to \infty, p/d \to \pi \in (0, 1], d/n \to \delta$.
- Let $\gamma := \pi\delta = \lim p/n$ and the resolvent moments: $\theta_j(\gamma, \lambda) := \int (x + \lambda)^{-j} dF_\gamma(x)$ where $F_\gamma(x)$ is the Marchenko-Pastur distribution with parameter γ.
- Let
$$\tilde{\lambda} := \lambda + \frac{1-\pi}{2\pi}\left[\lambda + 1 - \gamma + \sqrt{(\lambda+\gamma-1)^2 + 4\lambda}\right],$$
and $\tilde{\theta}_1 := \theta_1(\delta, \tilde{\lambda}), \tilde{\theta}_2 := \theta_2(\delta, \tilde{\lambda})$.

Theorem. Denoting s: features X; i: initialization W; l: label noise $\mathcal{E}$:

$$\lim_{d\to\infty} V_s = \alpha^2[1 - 2\tilde{\lambda}\tilde{\theta}_1 + \tilde{\lambda}^2\tilde{\theta}_2 - \pi^2(1-\lambda\theta_1)^2]$$
$$\lim_{d\to\infty} V_l = 0$$
$$\lim_{d\to\infty} V_i = \alpha^2\pi(1-\pi)(1-\lambda\theta_1)^2$$
$$\lim_{d\to\infty} V_{sl} = \sigma^2\delta(\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2)$$

$$\lim_{d\to\infty} V_{li} = 0$$
$$\begin{aligned}\lim_{d\to\infty} V_{si} = \alpha^2[&\pi(1 - 2\lambda\theta_1 + \lambda^2\theta_2 + (1-\pi)\delta(\theta_1 - \lambda\theta_2)) \\ &- \pi(1-\pi)(1-\lambda\theta_1)^2 - 1 + 2\tilde{\lambda}\tilde{\theta}_1 - \tilde{\lambda}^2\tilde{\theta}_2]\end{aligned}$$
$$\lim_{d\to\infty} V_{sli} = \sigma^2\delta[\pi(\theta_1 - \lambda\theta_2) - (\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2)].$$

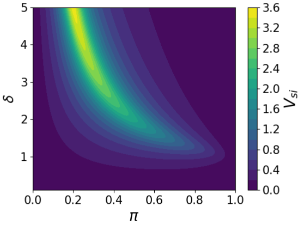
ANOVA for two-layer linear NN



Cumulative figure of the bias and variance components, as fn of $\delta = \lim d/n$. **Parameters**: signal strength $\alpha = 1$, noise level $\sigma = 0.3$, regularization parameter $\lambda = 0.01$, parametrization level $\pi = 0.8$.

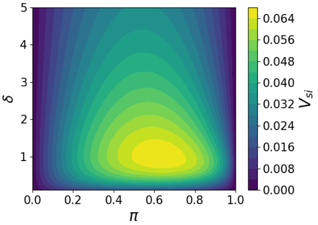
Interaction can dominate.

What is the effect of regularization?



V_{si}, $\lambda = 0.01$

optimal λ^* $\Longrightarrow$



V_{si}, $\lambda = \lambda^*$

- Large reduction in V_{si}
- V_{si}: The part of variance that can be reduced via ensembling over the sample X or initialization W.

Summary

- Calculus of Deterministic Equivalents: precise calculations of certain functionals of random matrices under mean-field asymptotics
- Compared to AMP: allows data distributions with more general covariance structure, but only for more specific trace functionals.
- Applications to distributed linear & ridge regression, random feature models.
- Other researchers' works in light of CDE: high-dimensional interpolation by Hastie et al 2019.

References I

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *arXiv preprint arXiv:2011.03321, NeurIPS 2020*, 2020.

David Barber, David Saad, and Peter Sollich. Finite-size effects and optimal test set size in linear perceptrons. *Journal of Physics A: Mathematical and General*, 28(5):1325, 1995.

Robert PW Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964. PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, 1995.

Lars Kai Hansen. Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6(3):393–396, 1993.

JA Hertz, A Krogh, and GI Thorbergsson. Phase transitions in simple learning. *Journal of Physics A: Mathematical and General*, 22(12): 2133, 1989.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

References II

Manfred Opper. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pages 922–925, 1995.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Sarunas Raudys and Robert PW Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392, 1998.