On Learning-Curve Monotonicity for Maximum Likelihood Estimators

Mark Sellke   Steven Yin

Abstract

The property of learning-curve monotonicity, highlighted in the recent papers [VML19, LVM19, VL22], describes algorithms which only improve in average performance given more data, for any underlying data distribution within a given family. We establish the first nontrivial monotonicity guarantees for the maximum likelihood estimator in a variety of well-specified parametric settings. For sequential prediction with log loss, we show monotonicity (in fact complete monotonicity) of the forward KL divergence for Gaussian vectors with unknown covariance and either known or unknown mean, as well as for Gamma variables with unknown scale parameter. The Gaussian setting was explicitly highlighted as open in the aforementioned works, even in dimension 1. Finally we observe that for reverse KL divergence, a folklore trick from e.g. [CT99] yields monotonicity for very general exponential families.

All results in this paper were derived by variants of GPT-5.2 Pro. Humans did not provide any proof strategies or intermediate arguments, but only prompted the model to continue developing additional results, and verified and transcribed its proofs.

Contents

1 Introduction

In statistics and machine learning, the classical notion of a *learning curve* describes how the performance of an algorithm evolves as it is exposed to more data. As highlighted in [VML19, LVM19, VL22], an appealing and intuitive property is *monotonicity* in the data: the average performance of a learning algorithm should only improve as more data becomes available. Although it is intuitively natural to expect such behavior, [VML19] constructed simple examples of non-monotone learning curves for mis-specified models, where the data generating distribution falls outside of the model class used for prediction (see also [LKB23]).

Does monotonicity hold for any commonly used estimators? We focus in particular on the maximum likelihood estimator under well-specification, addressing a question of [VL22]:

> *...we wonder whether maximum likelihood estimators for well-specified models behave monotonically. Likelihood estimation, being a century-old, classical technique, has been heavily studied, both theoretically and empirically. In much of the theory developed, the assumption that one is dealing with a correctly specified model is common, but we are not aware of any results that demonstrate that better models are obtained with more data.*

A particularly simple and fundamental instantiation of this general question was raised in [VML19]: fitting a single Gaussian. For IID samples from a Gaussian distribution with known unit variance but unknown mean, the n-sample MLE (i.e. the sample mean) differs from the true mean by a centered Gaussian with variance $\frac{1}{n}$; thus monotonicity holds in any reasonable sense. However monotonicity is much less clear for a Gaussian with known mean and *unknown* variance. Here we let $z_1, z_2, \dots$ be IID from a one-dimensional centered normal distribution $\mathcal{N}(0, v_*)$ with unknown variance $v_* > 0$. Then the maximum likelihood estimator for v_* is $\hat{v}_n = \frac{1}{n}\sum_{i=1}^n z_i^2$. We consider two risk measures for this approximation: the expected forward and reverse Kullback–Leibler (KL) divergence, given respectively by

$$\mathcal{E}_n = \mathbb{E}\big[D_{\mathrm{KL}}(\mathcal{N}(0, v_*), \mathcal{N}(0, \hat{v}_n))\big], \qquad \widetilde{\mathcal{E}}_n = \mathbb{E}\big[D_{\mathrm{KL}}(\mathcal{N}(0, \hat{v}_n), \mathcal{N}(0, v_*))\big].$$

In Example II therein, [VML19] demonstrated that a mis-specified analog of $\mathcal{E}_n$ (with z_i IID from a non-Gaussian distribution) could be non-monotone, and went on to ask:

> *...this raises the issue to what extent well-specified statistical models can actually be proven to behave monotonically. For instance, is Example II monotone if the problem is well-specified?*

[LVM19] similarly remarked that monotonicity is unclear for Gaussians when both the mean and variance are unknown. These works focused primarily on the forward KL risk $\mathcal{E}_n$, motivated by sequential prediction.

We establish monotonicity of the MLE in several well-specified settings. For forward KL divergence, we prove monotonicity for Gaussian vectors with unknown covariance and either known or unknown mean, as well as for Gamma distributions with unknown scale parameter but known shape (e.g. exponential random variables). In fact we show a much stronger property known as complete monotonicity, which implies that the marginal value of the n-th datapoint is not only positive but also strictly decreases with n. For reverse KL divergence, we observe that monotonicity holds in the extremely general setting of exponential families, a fact closely related to known results from e.g. [CT99, Chapter 2, Problem 34(b)].

Statement on AI Assistance

The proof of Theorem 1.1, which directly addresses the question of [VML19], was originally due to an unreleased prototype of a longer-thinking-time version of GPT-5.2 Pro. The model was asked to solve the 1-dimensional problem raised in [VML19], and came back to us with a correct proof in the more general setting of multivariate centered Gaussians. With this proof in context, the now-public version of GPT-5.2 Pro was then able to:

1. Transcribe the proof of Theorem 1.1 into Latex form.

2. Locate references for standard facts on Gamma variables and functions (the original AI output proved identities such as Equation (6) and Lemma 2.2 from scratch).

3. Extend the Gaussian results to unknown mean and reverse KL divergence, when we asked if such extensions were possible (see Theorem 1.2 and Propositions 1.6 and 1.7).

4. Extend the reverse KL results to Poisson and Binomial distributions using a convexity argument, when we asked if this was possible (GPT-5.2 Pro also pointed out that the expected forward KL divergences are always infinite in both cases).

5. Extend this convexity proof to general exponential families, when asked whether it could be pushed further (see Proposition 1.5).

6. Locate the closely related [CT99, Chapter 2, Problem 34(b)], when asked whether Proposition 1.5 was already known.

Having obtained these results, we wanted to test how the public GPT-5.2 Pro model would fare without being given the proof of Theorem 1.1 in the first step. Impressively, in its first attempt at the same initial query (itself an AI-generated restatement of the open problem from [VML19]), the public model also successfully proved Theorem 1.1. We then reproduced the other core results of this paper (Theorems 1.2 and 1.3, and Proposition 1.5) in our first attempt using only the following three follow-up prompts (without any retries or prompt rewriting):

- *how about in higher dimensions, when the entire covariance matrix is unknown?*

- *what about for exponential/gamma/poisson/binomial? does anything still work there?*

- *what about reverse KL? do things work there?*

On the other hand, our unreleased prototype also proved a much stronger *complete monotonicity* property of the forward KL divergence via an explicit Laplace transform representation. This extension was provided first (without our asking for it) in the originally requested setting of centered Gaussian variables in dimension 1. With this proof in context, GPT-5.2 Pro subsequently generalized it further to higher dimensional Gaussians, unknown means, and Gamma variables; see the discussion around Equation (3) and Section 4.

In summary, the human contributions to this paper (aside from the development of GPT-5.2 Pro and its internal variant) were as follows:

(a) Prompting the model to continue generalizing its results. We did not work out any of these extensions ourselves before prompting the model, but just asked simple and natural followup questions.

(b) Checking the proofs written by GPT-5.2 Pro. To our knowledge, the only mathematical mistake that needed fixing occurred in (7). GPT-5.2 Pro's initial proof flipped an inequality sign when bounding a sum by an integral, yielding an overly optimistic bound. However this mistake did not affect any downstream arguments.

(c) Editing and rearranging the writing for clarity and style, and writing the introduction.

Of course, all proofs, citations, and other claims were verified by the human authors, and we take full responsibility for their correctness.

1.1 Main Results

Let $Z_1, Z_2, \dots$ be IID from a distribution $P_* = P_{\theta_*}$ which belongs to a parametric family $\{P_\theta : \theta \in \Theta\}$ with densities $\{p_\theta\}_{\theta\in\Theta}$ (with respect to a fixed base measure). Given data $Z_{1:n} := (Z_1, \dots, Z_n)$, the MLE is any maximizer

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log p_\theta(Z_i).$$

We consider only settings where the MLE is almost surely unique, and study the expected accuracy of the fitted model $P_{\widehat{\theta}_n}$ as a function of n.

Learning-curve risks. To measure the estimation error between $\widehat{\theta}_n$ and θ_*, we consider both directions of the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(P\|Q) := \mathbb{E}_{X\sim P}\left[\log \frac{dP}{dQ}(X)\right].$$

Namely we define the *forward* and *reverse* KL learning curves of the MLE by

$$\mathcal{E}_n \;:=\; \mathbb{E}\big[D_{\mathrm{KL}}(P_* \,\|\, P_{\widehat{\theta}_n})\big]\,, \qquad \widetilde{\mathcal{E}}_n \;:=\; \mathbb{E}\big[D_{\mathrm{KL}}(P_{\widehat{\theta}_n} \,\|\, P_*)\big]\,,$$

where the expectation is over the training sample $Z_{1:n}$. We say the learning curve is monotone for forward KL divergence if $\mathcal{E}_{n+1} \le \mathcal{E}_n$ for all n where the former is finite, and strictly so if the inequality is always strict; similarly for $\widetilde{\mathcal{E}}_n$.

We note that in any parametric family with densities p_θ, the forward KL divergence is the excess expected log-loss of the prediction $P_{\widehat{\theta}_n}$ applied to IID data from P_*, relative to the true model, up to an additive constant independent of the estimator:

$$D_{\mathrm{KL}}(P_*\|P_{\widehat{\theta}_n}) = \mathbb{E}_{Z\sim P_*}[-\log p_{\widehat{\theta}_n}(Z)] \;-\; \mathbb{E}_{Z\sim P_*}[-\log p_{\theta_*}(Z)].$$

This sequential prediction perspective is the one taken in [VML19]. We will however work only with D_{KL}, which has the advantage of being invariant under reparametrization.

Gaussian models. Fix a dimension $d \ge 1$. In the centered multivariate Gaussian model, we observe $X_1 \dots, X_n \overset{IID}{\sim} \mathcal{N}(0, \Sigma_*)$ with unknown strictly positive definite covariance $\Sigma_* \in \mathbb{S}^d_{++}$. Given these observations, the MLE for the covariance Σ_* is given by

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

The corresponding forward-KL learning curve is

$$\mathcal{E}_{n,d}(\Sigma_*) := \mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathcal{N}(0,\Sigma_*) \,\|\, \mathcal{N}(0,\widehat{\Sigma}_n)\big)\Big].$$

In Section 3 we show the following monotonicity property of $\mathcal{E}_{n,d}$.

Theorem 1.1. *Fix $d \geq 1$ and $\Sigma_* \in \mathbb{S}^d_{++}$. If $n > d+1$, then the forward KL risk is independent of Σ_* and given by*

$$\mathcal{E}_{n,d}(\Sigma_*) = \frac{1}{2}\big(f_d(n) - d\big), \tag{1}$$

where ψ is the digamma function (see (5)) and

$$f_d(x) := \sum_{j=1}^{d} \psi\Big(\frac{x-j+1}{2}\Big) - d\log(x/2) + \frac{xd}{x-d-1}, \qquad x > d+1.$$

Moreover, $\mathcal{E}_{n,d}$ is data-monotone in the sense that

$$\mathcal{E}_{n+1,d}(\Sigma_*) \;<\; \mathcal{E}_{n,d}(\Sigma_*) \;<\; \infty.$$

If $n \leq d+1$, then $\mathcal{E}_{n,d}(\Sigma_*) = +\infty$.

We also consider a full Gaussian model, in which we observe $X_i \sim \mathcal{N}(\mu_*,\Sigma_*)$ with $(\mu_*,\Sigma_*) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$ with both μ_* and Σ_* unknown. The MLE is then given by:

$$\widehat{\mu}_n = \bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top.$$

We write $\mathcal{E}^{\mathrm{full}}_{n,d}(\mu_*,\Sigma_*)$ for the corresponding forward-KL learning curve of $(\widehat{\mu}_n, \widehat{\Sigma}_n)$ (see (11)). Here the finiteness threshold shifts to $n > d+2$. The next result is also shown in Section 3.

Theorem 1.2. *Fix $d \geq 1$, $\mu_* \in \mathbb{R}^d$, and $\Sigma_* \in \mathbb{S}^d_{++}$. If $n > d+2$, then the forward KL risk is independent of (μ_*,Σ_*) and given by*

$$\mathcal{E}^{\mathrm{full}}_{n,d}(\mu_*,\Sigma_*) = \frac{1}{2}\big(g_d(n) - d\big), \tag{2}$$

where ψ is the digamma function and

$$g_d(x) := \sum_{j=1}^{d} \psi\Big(\frac{x-j}{2}\Big) - d\log(x/2) + \frac{(x+1)d}{x-d-2}, \qquad x > d+2.$$

Moreover, $\mathcal{E}^{\mathrm{full}}_{n,d}$ is data-monotone in the sense that

$$\mathcal{E}^{\mathrm{full}}_{n+1,d}(\mu_*,\Sigma_*) < \mathcal{E}^{\mathrm{full}}_{n,d}(\mu_*,\Sigma_*) < \infty.$$

If $n \leq d+2$, then $\mathcal{E}^{\mathrm{full}}_{n,d}(\mu_*,\Sigma_*) = +\infty$.

In fact, both Gaussian settings can be reduced to the isotropic case $P_* = \mathcal{N}(0, I_d)$ via coordinate changes, which explains why the above formulas for $\mathcal{E}_{n,d}$ and $\widetilde{\mathcal{E}}_{n,d}$ have no dependence on μ_* or Σ_*. We derive the explicit polygamma function formulas in Theorems 1.1 and 1.2, and deduce monotonicity using derivative estimates for the trigamma function $\psi_1 = \psi'$.

Gamma-family scale estimation. A similar monotonicity extends to Gamma random variables. We fix a known shape parameter $\alpha > 0$ and seek to estimate the scale parameter θ_* from IID observations $X_i \sim \text{Gamma}(\alpha, \theta_*)$. The MLE is

$$\widehat{\theta}_n = \frac{1}{n\alpha} \sum_{i=1}^{n} X_i.$$

In Section 3.5 we prove the following.

Theorem 1.3. *If $n\alpha > 1$, then*

$$\mathbb{E}\Big[D_{\mathrm{KL}}(\text{Gamma}(\alpha, \theta_*) \,\|\, \text{Gamma}(\alpha, \widehat{\theta}_n))\Big] = h_\alpha(n) = \alpha\Big(\psi(n\alpha) - \log(n\alpha) + \frac{1}{n\alpha - 1}\Big),$$

where ψ is the digamma function, and this quantity is strictly decreasing in n. If $n\alpha \le 1$, the expectation is $+\infty$.

The special case $\alpha = 1$ of Theorem 1.3 recovers scale estimation for exponential random variables. Also notably, the case $2\alpha = d \in \mathbb{N}$ is equivalent to estimating the scale of a centered Gaussian which is *known to be isotropic*, i.e. of the form $\mathcal{N}(0, \theta_* I_d)$. Indeed here the squared norms $\|Z_i\|^2$ are IID $\text{Gamma}(d/2, 2\theta_*)$ and are sufficient statistics for θ_*.

Complete Monotonicity. An unreleased prototype of a longer-thinking-time version of GPT-5.2 Pro completed the proofs of monotonicity in Theorems 1.1, 1.2, and 1.3 using a different method. This approach rewrites the explicit functions above as Laplace transforms of positive kernels, i.e. in the form

$$L(x) = \int_0^\infty e^{-xt} d\mu(t) \tag{3}$$

where $\mu(t)$ is a positive finite measure on $[0, \infty)$ (for example a non-negative density). We detail this in Section 4, and obtain the following.

Theorem 1.4. *The functions f_d, g_d, h_α each have representations of the form (3), such that the integral converges absolutely on the respective half-lines $(d+1, \infty), (d+2, \infty), (1/\alpha, \infty)$. In particular, each $L \in \{f_d, g_d, h_\alpha\}$ satisfies the complete monotonicity inequalities*

$$(-1)^k L^{(k)}(x) > 0 \tag{4}$$

for all x in the domain of L and all $k \in \mathbb{N}$, where $L^{(k)}$ denotes the k-th derivative.

The inequalities (4) follow directly from (3) by differentiation under the integral sign. In fact the representation (3) famously characterizes all completely monotone functions;

see [Ber29, Mer14] or [Lax14, Chapter 14]. For integer n, the analogous discrete-difference properties follow by integration: for $k \in \mathbb{N}$ with all terms finite, one has for instance

$$\binom{k}{0}\mathcal{E}_{n,d} - \binom{k}{1}\mathcal{E}_{n+1,d} + \binom{k}{2}\mathcal{E}_{n+2,d} \cdots + (-1)^k \binom{k}{k}\mathcal{E}_{n+k,d} > 0.$$

In particular setting $k = 1$ recovers ordinary monotonicity. The case $k = 2$ (i.e. that $\mathcal{E}_{n,d} - 2\mathcal{E}_{n+1,d} + \mathcal{E}_{n+2,d} > 0$) also has a natural interpretation: although the added value of the n-th datapoint is strictly positive, it is also strictly decreasing with n.

Reverse KL Divergence. We also consider the same questions under reverse KL divergence

$$\widetilde{\mathcal{E}}_n = \mathbb{E}[D_{\mathrm{KL}}(P_{\widehat{\theta}_n} \| P_*)].$$

In fact, monotonicity of the reverse KL divergence for the MLE is known to hold in the case of discrete random variables; here the model class consists of all probability distributions on the support, which means the MLE is the empirical distribution. See for instance [CT99, Chapter 2, Problem 34(b)]. We observe (likely not for the first time) that the proof in the discrete case extends to the much more general setting of exponential families. Recall that an exponential family consists of densities of the form

$$p_\theta(x) = \exp(\langle \theta, T(x) \rangle - A(\theta)), \quad \theta \in \Theta \subseteq \mathbb{R}^k$$

with sufficient statistic T and log-partition function A. The family is regular if the mean map $\mu(\theta) = \mathbb{E}_\theta[T(X)] = \nabla A(\theta)$ is one-to-one and its range $\mathcal{M}$ is open. The convexity of the reverse KL divergence as a function of $\bar{T}_n := \frac{1}{n}\sum_{i=1}^n T(X_i)$ yields the following.

Proposition 1.5 (Reverse-KL data monotonicity in exponential families)**.** *Let $X_1, \ldots, X_n \sim p_{\theta_*}$ IID in a regular exponential family. Assume that for each n in a range of interest the MLE exists, lies in the interior, and hence satisfies $\widehat{\mu}_n := \nabla A(\widehat{\theta}_n) = \bar{T}_n$ almost surely. Then the expected reverse KL risk is nonincreasing:*

$$\mathbb{E}\Big[D_{\mathrm{KL}}(p_{\widehat{\theta}_{n+1}} \,\|\, p_{\theta_*})\Big] \le \mathbb{E}\big[D_{\mathrm{KL}}(p_{\widehat{\theta}_n} \,\|\, p_{\theta_*})\big] .$$

If in addition A^* is strictly convex on $\mathcal{M}$ (equivalently, the family is minimal) and $T(X)$ is non-degenerate under p_{θ_*}, then the inequality is strict.

The Gaussian settings are both special cases of the above, with $T(x) = xx^\top$ and $T(x) = (x, xx^\top)$ respectively. Below we provide alternate proofs of monotonicity for the reverse KL risk of Gaussian estimation, which were derived by GPT-5.2 Pro prior to Proposition 1.5 and again come with explicit formulas involving the digamma function. Of course, Proposition 1.5 encompasses other classical examples including Gamma, Poisson, and binomial random variables and many more.

Proposition 1.6. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, \Sigma_*)$ in $\mathbb{R}^d$, with MLE $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. If $n \ge d$, then the reverse KL risk is given by*

$$\mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathcal{N}(0, \widehat{\Sigma}_n) \,\|\, \mathcal{N}(0, \Sigma_*)\big)\Big] = \frac{1}{2}\Big(d \log(n/2) - \sum_{j=1}^d \psi\Big(\frac{n-j+1}{2}\Big)\Big),$$

and the expectation is strictly decreasing in n. If $n < d$, then the expectation is $+\infty$.

Proposition 1.7. *Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu_*, \Sigma_*)$ in $\mathbb{R}^d$, with MLEs*

$$\widehat{\mu}_n = \bar{X}, \qquad \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top.$$

If $n \geq d+1$, then the reverse KL risk is given by

$$\mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathcal{N}(\widehat{\mu}_n, \widehat{\Sigma}_n) \,\|\, \mathcal{N}(\mu_*, \Sigma_*)\big)\Big] = \frac{1}{2}\Big(d\log(n/2) - \sum_{j=1}^{d}\psi\Big(\frac{n-j}{2}\Big)\Big),$$

and the expectation is strictly decreasing in n. If $n \leq d$, then the expectation is $+\infty$.

1.2 Other Related Work

In the setting of sequential prediction, the realized risk R_n at time n is a function of the estimated parameter $\widehat{\theta}_n$ and the next sample z_n. Given a well-specified prior for θ_*, it is true in general that suitably defined Bayes-optimal algorithms have monotone learning curves, when averaged over the prior (see e.g. [VL22, Section 4.6]). For example, under log loss this is equivalent to the fact that posterior entropy decreases on average, a standard fact in information theory. See [HKOS91] for other results in this direction.

Learning curves also play a central role in the modern literature on large-scale machine learning, often via empirical scaling laws that relate test loss to dataset size and training compute [KMH$^+$20, HBM$^+$22]. Relatedly, hyperparameter transfer methods aim to predict or accelerate progress along such curves across scales [YHB$^+$21]. The notable double descent phenomenon [BHMM19, BHX20, NKB$^+$21, HMRT22, MM22, LW21] is a surprising non-monotonicity property of the test loss relative to model size. These works document striking regularities in overparameterized regimes, but they are largely orthogonal to our focus of well-specified parametric models. On the other hand, our Gaussian example does include high-dimensional linear regression with Gaussian covariates and errors, implying that the log-loss of the maximum likelihood linear model is data monotone, even when the noise level is one of the quantities to be estimated.

Classical theoretical analyses of learning curves in Bayesian/statistical-mechanics settings go back at least to [PC87, SST92, OH91, AFS92, Ama93], with additional viewpoints from PAC/VC-style estimation [TT93, HKOS91]. The book [DGL97] asked whether monotone Bayes-consistent algorithms exist; this was resolved positively in great generality by [Pes22], and in a black-box way by [BDK$^+$22]. Our results show that the MLE suffices for monotonicity in several natural examples.

2 Gamma functions and distributions

The Gamma function and associated objects will play a key role in our proofs. Recall that

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt, \qquad z > 0.$$

The digamma and trigamma functions are defined by logarithmic differentiation:

$$\psi(z) = \frac{d}{dz}\log\Gamma(z), \qquad \psi_1(z) = \psi'(z) = \frac{d^2}{dz^2}\log\Gamma(z). \tag{5}$$

We use a standard series identity for ψ_1 (see e.g. [BE53, Section 1.16, Equation (9)]):

$$\psi_1(z) = \sum_{k=0}^{\infty} \frac{1}{(z+k)^2}, \qquad \forall z > 0. \tag{6}$$

This identity yields the following estimates which will be key in verifying monotonicity.

Lemma 2.1. *Assuming $t > 0$ in the first inequality and $t > 1$ in the second, we have*

$$\psi_1(t) < \frac{t+1}{t^2} < \frac{1}{t-1}. \tag{7}$$

For $t > 0$ we also have the lower bound:

$$\psi_1(t) > \frac{1}{t}. \tag{8}$$

Proof. The latter estimate in (7) is clear. For the former, we have from (6) that

$$\psi_1(t) < \frac{1}{t^2} + \sum_{m=1}^{\infty} \frac{1}{(t+m)(t+m-1)} = \frac{1}{t^2} + \frac{1}{t} = \frac{t+1}{t^2}.$$

For (8), we have

$$\psi_1(t) = \sum_{k=0}^{\infty} \frac{1}{(t+k)^2} > \int_0^{\infty} \frac{ds}{(t+s)^2} = \frac{1}{t}. \qquad \square$$

Next we recall the Gamma distribution: for a shape parameter $\alpha > 0$ and scale parameter $\theta > 0$, we write $V \sim \mathrm{Gamma}(\alpha, \theta)$ if V has density

$$f(v) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} v^{\alpha-1} e^{-v/\theta}, \qquad v > 0. \tag{9}$$

An important special case is the chi-squared distribution $\chi^2_\nu = \mathrm{Gamma}(\nu/2, 2)$ for $\nu > 0$. Recall that χ^2_ν is the law of $Z_1^2 + \cdots + Z_\nu^2$ for IID standard Gaussian Z_i when $\nu \in \mathbb{N}$.

Lemma 2.2 ([KBJ19, Equation (17.29)])**.** *If $V \sim \mathrm{Gamma}(\alpha, \theta)$, then*

$$\mathbb{E}[\log V] = \psi(\alpha) + \log\theta.$$

In particular, if $U \sim \chi^2_\nu = \mathrm{Gamma}(\nu/2, 2)$, then

$$\mathbb{E}[\log U] = \psi\left(\frac{\nu}{2}\right) + \log 2.$$

Lemma 2.3 (Reciprocal moment of a Gamma random variable)**.** *If $V \sim \mathrm{Gamma}(\alpha, \theta)$ and $\alpha > 1$, then*

$$\mathbb{E}[V^{-1}] = \frac{1}{\theta(\alpha - 1)}.$$

If $\alpha \leq 1$, then $\mathbb{E}[V^{-1}] = +\infty$. In particular, if $U \sim \chi^2_\nu$ with $\nu > 2$, then

$$\mathbb{E}[U^{-1}] = \frac{1}{\nu - 2}$$

and $\mathbb{E}[U^{-1}] = \infty$ if $\nu \leq 2$.

Proof. Direct integration gives, for $\alpha > 1$,

$$\begin{aligned}\mathbb{E}[V^{-1}] = \int_0^\infty s^{-1} \frac{1}{\Gamma(\alpha)\theta^\alpha} s^{\alpha-1} e^{-s/\theta}\, ds &= \frac{1}{\Gamma(\alpha)\theta^\alpha} \int_0^\infty s^{\alpha-2} e^{-s/\theta}\, ds \\ &= \frac{\Gamma(\alpha-1)\theta^{\alpha-1}}{\Gamma(\alpha)\theta^\alpha} = \frac{1}{\theta(\alpha-1)}.\end{aligned}$$

If $\alpha \leq 1$, the integral diverges at 0. □

3 Gaussian and Gamma estimation under forward KL

3.1 Coordinate Change Reductions for Gaussians

Here we reduce both the known and unknown mean cases of Gaussian estimation to $P_* = \mathcal{N}(0, I_d)$ by coordinate change.

Lemma 3.1 (Coordinate change for known mean)**.** *Let $Y_i := \Sigma_*^{-1/2} X_i$, so $Y_i \sim \mathcal{N}(0, I_d)$, and define*

$$\widehat{S}_n := \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\top.$$

Then $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top = \Sigma_*^{1/2} \widehat{S}_n \Sigma_*^{1/2}$ and

$$D_{\mathrm{KL}}\big(\mathcal{N}(0, \Sigma_*) \,\|\, \mathcal{N}(0, \widehat{\Sigma}_n)\big) = \frac{1}{2}\Big(\log\det \widehat{S}_n + \mathrm{tr}(\widehat{S}_n^{-1}) - d\Big)$$

and so

$$\mathcal{E}_{n,d}(\Sigma_*) = \frac{1}{2}\Big(\mathbb{E}[\log\det \widehat{S}_n] + \mathbb{E}[\mathrm{tr}(\widehat{S}_n^{-1})] - d\Big). \tag{10}$$

Proof. The covariance identity is immediate from $X_i = \Sigma_*^{1/2} Y_i$. Moreover $\log\det \widehat{\Sigma}_n = \log\det \Sigma_* + \log\det \widehat{S}_n$. For the trace term,

$$\mathrm{tr}(\widehat{\Sigma}_n^{-1}\Sigma_*) = \mathrm{tr}\left((\Sigma_*^{1/2}\widehat{S}_n\Sigma_*^{1/2})^{-1}\Sigma_*\right) = \mathrm{tr}\left(\Sigma_*^{-1/2}\widehat{S}_n^{-1}\Sigma_*^{-1/2}\Sigma_*\right) = \mathrm{tr}(\widehat{S}_n^{-1}),$$

by cyclicity of trace. Substituting into $D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma_*)\|\mathcal{N}(0, \widehat{\Sigma}_n))$ and taking expectations yields (10). □

We omit the very similar proof for the unknown mean case, stated below.

Lemma 3.2 (Coordinate change for the full model)**.** *Let* $Y_i := \Sigma_*^{-1/2}(X_i - \mu_*)$ *and* $\bar{Y} := \frac{1}{n}\sum_{i=1}^n Y_i$. *We now let*

$$\widehat{S}_n := \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top.$$

Then $Y_i \sim \mathcal{N}(0, I_d)$ *and the MLE is given by*

$$\widehat{\mu}_n - \mu_* = \Sigma_*^{1/2}\bar{Y}, \qquad \widehat{\Sigma}_n = \Sigma_*^{1/2}\widehat{S}_n\Sigma_*^{1/2}.$$

Moreover,

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu_*, \Sigma_*) \,\|\, \mathcal{N}(\widehat{\mu}_n, \widehat{\Sigma}_n)\big) = \frac{1}{2}\Big(\log\det \widehat{S}_n + \mathrm{tr}(\widehat{S}_n^{-1}) + \bar{Y}^\top \widehat{S}_n^{-1}\bar{Y} - d\Big)$$

and so

$$\mathcal{E}_{n,d}^{\mathrm{full}}(\mu_*, \Sigma_*) = \frac{1}{2}\Big(\mathbb{E}[\log\det \widehat{S}_n] + \mathbb{E}[\mathrm{tr}(\widehat{S}_n^{-1})] + \mathbb{E}[\bar{Y}^\top \widehat{S}_n^{-1}\bar{Y}] - d\Big). \tag{11}$$

3.2 Wishart Matrices

Let $Y_1, \ldots, Y_n \sim \mathcal{N}(0, I_d)$ IID and define the Wishart matrix

$$W := \sum_{i=1}^n Y_i Y_i^\top \sim \mathrm{Wishart}_d(n, I_d), \qquad \widehat{S}_n = \frac{1}{n}W.$$

We rely on the following well-known consequence of Gram–Schmidt orthogonalization.

Lemma 3.3 ([And03, Lemma 7.2.1])**.** *Let* $W \sim \mathrm{Wishart}_d(n, I_d)$ *with* $n \geq d$. *Then*

$$\det W \;\overset{d}{=}\; \prod_{j=1}^d \chi^2_{n-j+1},$$

where the chi-squared variables on the right-hand side are independent.

Proposition 3.4. *Let* $W \sim \mathrm{Wishart}_d(n, I_d)$ *with* $n \geq d$. *Then*

$$\mathbb{E}[\log\det W] = d\log 2 + \sum_{j=1}^d \psi\Big(\frac{n-j+1}{2}\Big).$$

Proof. By Lemma 3.3, $\log\det W \overset{d}{=} \sum_{j=1}^d \log\chi^2_{n-j+1}$. Apply Lemma 2.2 termwise. □

Lemma 3.5. *Let* $W \sim \mathrm{Wishart}_d(n, I_d)$ *with* $n \geq d$. *Then*

$$(W^{-1})_{11} \;\overset{d}{=}\; \frac{1}{\chi^2_{n-d+1}}.$$

Proof. Let Z be the $n \times d$ data matrix with rows $Y_1^\top, \dots, Y_n^\top$, so that $W = Z^\top Z$. Write

$$Z = [z_1 \; Z_2],$$

where $z_1 \in \mathbb{R}^n$ is the first column and $Z_2 \in \mathbb{R}^{n\times(d-1)}$ contains the remaining columns. Then

$$W = Z^\top Z = \begin{pmatrix} z_1^\top z_1 & z_1^\top Z_2 \\ Z_2^\top z_1 & Z_2^\top Z_2 \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12}^\top \\ w_{12} & W_{22} \end{pmatrix}.$$

The formula for the inverse of a block matrix is

$$(W^{-1})_{11} = \frac{1}{w_{11} - w_{12}^\top W_{22}^{-1} w_{12}}.$$

It is well-known (see e.g. [Mui09, Theorem 3.2.10]) that the denominator is a chi-squared random variable with the claimed number of degrees of freedom, as desired. Explicitly, we may write

$$w_{11} - w_{12}^\top W_{22}^{-1} w_{12} = z_1^\top \left(I_n - Z_2 (Z_2^\top Z_2)^{-1} Z_2^\top \right) z_1.$$

The matrix

$$Q = I_n - Z_2 (Z_2^\top Z_2)^{-1} Z_2^\top$$

is the orthogonal projector onto $\mathrm{col}(Z_2)^\perp$ and has rank $n - d + 1$ almost surely. Since $z_1 \sim \mathcal{N}(0, I_n)$ and is independent of Z_2, it follows that $z_1^\top Q z_1 \sim \chi^2_{n-d+1}$ is equal in distribution to the sum of squares of $n - d + 1$ IID standard Gaussians, as claimed. □

Proposition 3.6. *Let $W \sim \mathrm{Wishart}_d(n, I_d)$ with $n \geq d$. If $n \leq d + 1$ then $\mathbb{E}[(W^{-1})_{11}] = \mathbb{E}[\mathrm{tr}(W^{-1})] = \infty$. If $n > d + 1$ then*

$$\mathbb{E}[W^{-1}] = \frac{I_d}{n - d - 1}, \qquad \mathbb{E}[\mathrm{tr}(W^{-1})] = \frac{d}{n - d - 1}.$$

Proof. Lemma 3.5 and Lemma 2.3 give $\mathbb{E}[(W^{-1})_{11}] = 1/(n - d - 1)$ for $n > d + 1$ and $+\infty$ otherwise. Positive definiteness implies the off-diagonal entries have finite expectation when the diagonal entries do. Orthogonal invariance of W implies $\mathbb{E}[W^{-1}] \propto I_d$ when the expectation exists. Combining completes the proof. □

3.3 Known mean: closed form and monotonicity

Proof of Theorem 1.1. We first show (1). This follows by combining Lemma 3.1 with Proposition 3.4 and Proposition 3.6, using $\widehat{S}_n = W/n$, which yields

$$\log\det \widehat{S}_n = \log\det W - d \log n, \quad \mathrm{tr}(\widehat{S}_n^{-1}) = n\,\mathrm{tr}(W^{-1}).$$

To complete the proof, it suffices to show $f_d'(x) < 0$ for all $x > d + 1$. Differentiating yields:

$$f_d'(x) = \frac{1}{2} \sum_{j=1}^{d} \psi_1\Big(\frac{x - j + 1}{2}\Big) - \frac{d}{x} - \frac{d(d+1)}{(x - d - 1)^2}.$$

By (7), we have $\frac{1}{2}\psi_1((x-j+1)/2) < \frac{1}{x-j-1}$, hence

$$f_d'(x) < \sum_{j=1}^{d} \frac{1}{x-j-1} - \frac{d}{x} - \frac{d(d+1)}{(x-d-1)^2} \le d\Big(\frac{1}{x-d-1} - \frac{1}{x} - \frac{d+1}{(x-d-1)^2}\Big)$$
$$= \frac{-d(d+1)^2}{x(x-d-1)^2} < 0. \qquad \square$$

3.4 Unknown mean: closed form and monotonicity

For a Gaussian with unknown mean, recentering around the empirical sample mean is essentially equivalent to removing a single datapoint. We detail this below.

Lemma 3.7. *Let $Y_1, \dots, Y_n \sim \mathcal{N}(0, I_d)$ be IID and $\bar{Y} = \frac{1}{n}\sum_i Y_i$. Define*

$$W_c := \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^\top.$$

Then $W_c \sim \mathrm{Wishart}_d(n-1, I_d)$ and W_c is independent of $\bar{Y}$.

Proof. Let $Z \in \mathbb{R}^{n\times d}$ be the data matrix with rows $Y_1^\top, \dots, Y_n^\top$. Let $\mathbf{1}_n \in \mathbb{R}^n$ be the all-ones vector and set $q_1 := \mathbf{1}_n/\sqrt{n}$. Extend q_1 to an orthonormal basis $q_1, \dots, q_n$ and let $Q \in \mathbb{R}^{n\times n}$ be the orthogonal matrix with rows $q_1^\top, \dots, q_n^\top$. Since Z is Gaussian with IID $\mathcal{N}(0,1)$ entries, QZ has the same distribution and has independent rows. Writing

$$QZ = \begin{pmatrix} \sqrt{n}\,\bar{Y}^\top \\ \widetilde{Z} \end{pmatrix},$$

the matrix $\widetilde{Z} \in \mathbb{R}^{(n-1)\times d}$ has IID $\mathcal{N}(0,1)$ entries and is independent of $\bar{Y}$. Moreover,

$$W_c = Z^\top \Big(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\Big) Z = Z^\top Q^\top \begin{pmatrix} 0 & 0 \\ 0 & I_{n-1} \end{pmatrix} QZ = \widetilde{Z}^\top \widetilde{Z},$$

so $W_c \sim \mathrm{Wishart}_d(n-1, I_d)$ and is independent of $\bar{Y}$. $\square$

Proof of Theorem 1.2. We start with (2). By Lemma 3.2 we may assume $Y_i \sim \mathcal{N}(0, I_d)$. Let W_c be as in Lemma 3.7; then $\widehat{S}_n = W_c/n$. The log-determinant term is given by Proposition 3.4 with $n-1$ in place of n. The inverse-trace term is given by Proposition 3.6 with the same substitution. For the mean term, independence gives

$$\mathbb{E}[\bar{Y}^\top \widehat{S}_n^{-1} \bar{Y}] = \mathrm{tr}\Big(\mathbb{E}[\widehat{S}_n^{-1}]\,\mathbb{E}[\bar{Y}\bar{Y}^\top]\Big) = \mathrm{tr}\Big(\frac{n}{n-d-2} I_d \cdot \frac{1}{n} I_d\Big) = \frac{d}{n-d-2}.$$

As in the centered case, it remains to show $g_d'(x) < 0$ for all $x > d+2$. Differentiating:

$$g_d'(x) = \frac{1}{2}\sum_{j=1}^{d} \psi_1\Big(\frac{x-j}{2}\Big) - \frac{d}{x} - \frac{d(d+3)}{(x-d-2)^2}.$$

By (7), $\frac{1}{2}\psi_1((x-j)/2) < \frac{1}{x-j-2}$, hence

$$g_d'(x) < \Big(\sum_{j=1}^{d} \frac{1}{x-j-2}\Big) - \frac{d}{x} - \frac{d(d+3)}{(x-d-2)^2} \le d\Big(\frac{1}{x-d-2} - \frac{1}{x} - \frac{d+3}{(x-d-2)^2}\Big)$$
$$= \frac{-d\big(x+(d+2)^2\big)}{x(x-d-2)^2} < 0. \qquad \square$$

3.5 Gamma Distributions under Forward KL

Fix $\alpha > 0$ and let $X_i \sim \mathrm{Gamma}(\alpha, \theta_*)$ with unknown scale $\theta_* > 0$ and known shape α. The MLE is

$$\widehat{\theta}_n = \frac{1}{n\alpha}\sum_{i=1}^{n} X_i. \tag{12}$$

Indeed, the log-likelihood as a function of θ is

$$\ell(\theta) = -n\alpha \log\theta - \theta^{-1}\sum_i X_i + C$$

where C does not depend on θ; setting $\ell'(\theta) = 0$ yields (12).

Proof of Theorem 1.3. For fixed shape α, one computes

$$D_{\mathrm{KL}}(\mathrm{Gamma}(\alpha,\theta_*) \,\|\, \mathrm{Gamma}(\alpha,\theta)) = \alpha\Big(\log\frac{\theta}{\theta_*} + \frac{\theta_*}{\theta} - 1\Big).$$

Indeed, the log-density ratio for $X \sim \mathrm{Gamma}(\alpha,\theta_*)$ is

$$\log\frac{f_{\theta_*}(X)}{f_\theta(X)} = \Big(-\alpha\log\theta_* - \frac{X}{\theta_*}\Big) - \Big(-\alpha\log\theta - \frac{X}{\theta}\Big) = \alpha\log\frac{\theta}{\theta_*} + X\Big(\frac{1}{\theta} - \frac{1}{\theta_*}\Big).$$

Taking expectation and using $\mathbb{E}_{\theta_*}[X] = \alpha\theta_*$ gives

$$D_{\mathrm{KL}}(\mathrm{Gamma}(\alpha,\theta_*) \,\|\, \mathrm{Gamma}(\alpha,\theta)) = \alpha\log\frac{\theta}{\theta_*} + \alpha\theta_*\Big(\frac{1}{\theta} - \frac{1}{\theta_*}\Big) = \alpha\Big(\log\frac{\theta}{\theta_*} + \frac{\theta_*}{\theta} - 1\Big).$$

Let $S = \sum_i X_i \sim \mathrm{Gamma}(n\alpha, \theta_*)$, so

$$\widehat{\theta}_n = \frac{S}{n\alpha} \sim \mathrm{Gamma}\Big(n\alpha, \frac{\theta_*}{n\alpha}\Big).$$

Then Lemma 2.2 gives $\mathbb{E}[\log(\widehat{\theta}_n/\theta_*)] = \psi(n\alpha) - \log(n\alpha)$. Lemma 2.3 gives (for $n\alpha > 1$)

$$\mathbb{E}\Big[\frac{\theta_*}{\widehat{\theta}_n}\Big] = \theta_*\,\mathbb{E}[\widehat{\theta}_n^{-1}] = \frac{\theta_* n\alpha}{\theta_*(n\alpha-1)} = \frac{n\alpha}{n\alpha-1}.$$

Monotonicity follows by differentiating in $t = n\alpha$ and using (7). Concretely, define

$$h(t) := \alpha\Big(\psi(t) - \log t + \frac{1}{t-1}\Big), \quad \forall t > 1.$$

Then $h(n\alpha)$ is the displayed risk, and

$$h'(t) = \alpha\Big(\psi_1(t) - \frac{1}{t} - \frac{1}{(t-1)^2}\Big).$$

By (7), $\psi_1(t) < \frac{t+1}{t^2} < \frac{1}{t} + \frac{1}{(t-1)^2}$ for $t > 1$; hence $h'(t) < 0$, implying the result. $\square$

4 Complete monotonicity of Forward KL Divergence

Here we finish the proofs of Theorems 1.1, 1.2, and 1.3 in a different way. Namely we represent the relevant functions as Laplace transforms of positive measures, i.e. in the form

$$L(x) = \int_0^\infty e^{-xt} d\mu(t).$$

This representation implies the much stronger property of *complete monotonicity*: such a function L satisfies $(-1)^k L^{(k)}(x) \geq 0$ where $L^{(k)}$ is the k-th derivative.

We will use the standard Laplace transform identities:

$$\frac{1}{x} = \int_0^\infty e^{-xt}\, dt; \tag{13}$$

$$\frac{1}{(x-w)^2} = \int_0^\infty t\, e^{-(x-w)t}\, dt = \int_0^\infty t\, e^{-xt} e^{wt}\, dt, \quad \forall x > w. \tag{14}$$

By expanding the series (6) for ψ_1 we also obtain the Laplace transform representation of the trigamma function.

Lemma 4.1. *For every $u > 0$,*

$$\psi_1(u) = \int_0^\infty \frac{t\, e^{-ut}}{1 - e^{-t}}\, dt. \tag{15}$$

Proof. Starting from (6) and the identity $\frac{1}{a^2} = \int_0^\infty t e^{-at}\, dt$ for $a > 0$, we obtain

$$\psi_1(u) = \sum_{k=0}^\infty \int_0^\infty t e^{-(u+k)t}\, dt = \int_0^\infty t e^{-ut} \sum_{k=0}^\infty e^{-kt}\, dt = \int_0^\infty \frac{t\, e^{-ut}}{1 - e^{-t}}\, dt,$$

where Tonelli's theorem justifies exchanging sum and integral. □

We now prove Theorem 1.4, proceeding separately for each of the three cases. In fact we will show the functions $-f_d', -g_d', -h_\alpha'$ are Laplace transforms of positive kernels as in (3). This implies they are completely monotone, hence so are f_d, g_d, h_α. (By definition of complete monotonicity it remains to verify their non-negativity; this follows from non-negativity of KL divergence.) This yields Theorem 1.4 by again applying the equivalence between (3) and completely monotone functions in the opposite direction (see e.g. [Lax14, Chapter 14]).

Proof of Theorem 1.4 for f_d. Fix $x > d + 1$. Using (13), (14), (15) with $w = d + 1$ and the change of variables $s \mapsto 2t$, we compute

$$\frac{1}{2}\psi_1\Big(\frac{x-j+1}{2}\Big) = \frac{1}{2}\int_0^\infty \frac{s\, e^{-(x-j+1)s/2}}{1 - e^{-s}}\, ds = \int_0^\infty \frac{2t\, e^{-(x-j+1)t}}{1 - e^{-2t}}\, dt.$$

Summing over $j = 1, \dots, d$ and using $\sum_{j=1}^d e^{(j-1)t} = \frac{e^{dt}-1}{e^t-1}$ gives

$$\frac{1}{2}\sum_{j=1}^d \psi_1\Big(\frac{x-j+1}{2}\Big) = \int_0^\infty e^{-xt}\, \frac{2t}{1-e^{-2t}} \cdot \frac{e^{dt}-1}{e^t - 1}\, dt.$$

Substituting these representations into $-f_d'(x) = \frac{d}{x} + \frac{d(d+1)}{(x-d-1)^2} - \frac{1}{2}\sum_j \psi_1(\frac{x-j+1}{2})$ yields

$$-f_d'(x) = \int_0^\infty e^{-xt}\, B_d(t)\, dt, \tag{16}$$

where the kernel is

$$B_d(t) := d + d(d+1)t\, e^{(d+1)t} - \frac{2t}{1-e^{-2t}} \cdot \frac{e^{dt}-1}{e^t-1}, \qquad t > 0. \tag{17}$$

The integral in (16) is absolutely convergent for $x > d+1$: as $t \downarrow 0$ one has $B_d(t) = O(t)$ (hence no singularity at 0), while as $t \to \infty$, $B_d(t) = O(te^{(d+1)t})$ and e^{-xt} provides exponential decay since $x > d+1$.

It remains to show $B_d(t) > 0$ for all $t > 0$. First note

$$\frac{2t}{1-e^{-2t}} = \frac{te^t}{\sinh t} \le e^t \qquad (t > 0),$$

since $\sinh t \ge t$ for $t \ge 0$. Hence

$$B_d(t) \;\ge\; d + d(d+1)t\, e^{(d+1)t} - e^t \cdot \frac{e^{dt}-1}{e^t-1}.$$

Using $e^t \cdot \frac{e^{dt}-1}{e^t-1} = \sum_{k=1}^d e^{kt}$, we obtain the simpler lower bound

$$B_d(t) \;\ge\; F_d(t) := d + d(d+1)t\, e^{(d+1)t} - \sum_{k=1}^d e^{kt}.$$

Now $F_d(0) = 0$, and differentiating gives

$$F_d'(t) = d(d+1)e^{(d+1)t}\big(1 + (d+1)t\big) - \sum_{k=1}^d ke^{kt}.$$

Since $e^{kt} \le e^{dt}$ for $1 \le k \le d$, we have

$$\sum_{k=1}^d ke^{kt} \le e^{dt}\sum_{k=1}^d k = \frac{d(d+1)}{2}\, e^{dt}.$$

Therefore,

$$F_d'(t) \ge d(d+1)e^{dt}\Big(e^t(1 + (d+1)t) - \frac{1}{2}\Big).$$

The right-hand side is strictly positive for $t > 0$ because $e^t(1+(d+1)t) > 1$. Thus $F_d(t) > 0$ for all $t > 0$, and hence $B_d(t) \ge F_d(t) > 0$ for all $t > 0$. □

Proof of Theorem 1.4 for g_d. Recall that for $x > d+2$ we have

$$g_d'(x) = \frac{1}{2}\sum_{j=1}^d \psi_1\Big(\frac{x-j}{2}\Big) - \frac{d}{x} - \frac{d(d+3)}{(x-d-2)^2}.$$

Fix $x > d+2$. Using (13), (14), (15) with $w = d+2$ and the change of variables $s \mapsto 2t$, we compute

$$\frac{1}{2}\psi_1\Big(\frac{x-j}{2}\Big) = \frac{1}{2}\int_0^\infty \frac{s\,e^{-(x-j)s/2}}{1-e^{-s}}\,ds = \int_0^\infty \frac{2t\,e^{-(x-j)t}}{1-e^{-2t}}\,dt.$$

Summing over $j = 1, \dots, d$ yields

$$\frac{1}{2}\sum_{j=1}^{d}\psi_1\Big(\frac{x-j}{2}\Big) = \int_0^\infty e^{-xt}\,\frac{2t}{1-e^{-2t}}\cdot\sum_{j=1}^{d} e^{jt}\,dt = \int_0^\infty e^{-xt}\,\frac{2t}{1-e^{-2t}}\cdot\frac{e^{(d+1)t}-e^t}{e^t-1}\,dt.$$

Substituting into

$$-g_d'(x) = \frac{d}{x} + \frac{d(d+3)}{(x-d-2)^2} - \frac{1}{2}\sum_{j=1}^{d}\psi_1\Big(\frac{x-j}{2}\Big)$$

gives the Laplace transform representation

$$-g_d'(x) = \int_0^\infty e^{-xt}\,\widetilde{B}_d(t)\,dt, \tag{18}$$

where the kernel is

$$\widetilde{B}_d(t) := d+d(d+3)t\,e^{(d+2)t} - \frac{2t}{1-e^{-2t}}\cdot\sum_{j=1}^{d} e^{jt} \;=\; d+d(d+3)t\,e^{(d+2)t} - \frac{2t}{1-e^{-2t}}\cdot\frac{e^{(d+1)t}-e^t}{e^t-1}.$$

The integrand in (18) is again absolutely convergent for $x > d+2$.

We now show $\widetilde{B}_d(t) > 0$ for all $t > 0$. Using again that $\frac{2t}{1-e^{-2t}} \le e^t$ for $t > 0$ we find

$$\widetilde{B}_d(t) \ge \widetilde{F}_d(t) := d + d(d+3)t\,e^{(d+2)t} - e^t\sum_{j=1}^{d} e^{jt} = d + d(d+3)t\,e^{(d+2)t} - \sum_{k=2}^{d+1} e^{kt}.$$

Then $\widetilde{F}_d(0) = 0$, and differentiating gives

$$\widetilde{F}_d'(t) = d(d+3)e^{(d+2)t}\big(1+(d+2)t\big) - \sum_{k=2}^{d+1} ke^{kt}.$$

Since $e^{kt} \le e^{(d+1)t}$ for $2 \le k \le d+1$ and

$$\sum_{k=2}^{d+1} k = \frac{(d+1)(d+2)}{2} - 1 = \frac{d(d+3)}{2},$$

we have

$$\sum_{k=2}^{d+1} ke^{kt} \le \frac{d(d+3)}{2}\,e^{(d+1)t}.$$

Therefore,

$$\widetilde{F}_d'(t) \ge d(d+3)e^{(d+1)t}\Big(e^t(1+(d+2)t) - \frac{1}{2}\Big) > 0 \qquad (t > 0),$$

because $e^t(1+(d+2)t) \ge 1$ for all $t \ge 0$ and is strictly > 1 for $t > 0$. Thus $\widetilde{F}_d(t) > 0$ for all $t > 0$, and hence $\widetilde{B}_d(t) \ge \widetilde{F}_d(t) > 0$ for all $t > 0$. □

Proof of Theorem 1.4 for h_α. For $t := n\alpha$, the expected forward KL risk equals

$$\mathcal{R}_{n,\alpha}(\theta_*) := \mathbb{E}\Big[D_{\mathrm{KL}}(\mathrm{Gamma}(\alpha,\theta_*) \,\|\, \mathrm{Gamma}(\alpha,\widehat{\theta}_n))\Big] = \alpha\, h(t); \qquad h(t) := \psi(t) - \log t + \frac{1}{t-1},$$

whenever $t > 1$ (and $\mathcal{R}_{n,\alpha} = +\infty$ for $t \le 1$).

Differentiating,

$$h'(t) = \psi_1(t) - \frac{1}{t} - \frac{1}{(t-1)^2}.$$

Using (13), (14), (15) with $w = 1$ we obtain

$$-h'(t) = \int_0^\infty e^{-ts}\, \breve{B}(s)\, ds; \qquad \breve{B}(s) := 1 + se^s - \frac{s}{1-e^{-s}}. \tag{19}$$

To see that $\breve{B}(s) > 0$ for all $s > 0$, set $u = e^s > 1$ and set

$$\breve{F}(u) := (u-1)\breve{B}(s) = (u-1) + u(u-2)\log u.$$

Then one has

$$\breve{F}'(u) = (u-1)(2\log u + 1) > 0, \quad \forall u > 1.$$

Meanwhile $\breve{F}(1) = 0$, and so $\breve{F}(u) > 0$ for all $u > 1$. This implies $\breve{B}(s) > 0$ for all $s > 0$ as desired. □

5 Reverse KL monotonicity in exponential families

In this section we prove Proposition 1.5 on reverse KL monotonicity in general exponential families, and then present the explicit Gaussian formulas in Proposition 1.6 and 1.7.

To generalize the monotonicity of reverse KL divergence, we take advantage of the following domination in the convex order.

Lemma 5.1. *Let $Z_1, Z_2, \dots$ be IID in a finite-dimensional real vector space V, and let $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$. Let $\phi : V \to (-\infty, +\infty]$ be convex and assume $\mathbb{E}[\phi(\bar{Z}_{n_0})] < \infty$. Then*

$$\mathbb{E}[\phi(\bar{Z}_{n+1})] \le \mathbb{E}[\phi(\bar{Z}_n)], \quad \forall n \ge n_0.$$

If ϕ is strictly convex on a convex set supporting $\bar{Z}_n$ and Z_1 is non-degenerate, then the inequality is strict.

Proof. Let

$$\bar{Z}_n^{(-i)} := \frac{1}{n} \sum_{\substack{1 \le j \le n+1 \\ j \ne i}} Z_j$$

be the leave-one-out mean. Let I be uniform on $\{1, \dots, n+1\}$, independent of all Z_j. Then $\mathbb{E}[\bar{Z}_n^{(-I)} \mid Z_1, \dots, Z_{n+1}] = \bar{Z}_{n+1}$. Jensen gives

$$\phi(\bar{Z}_{n+1}) \le \mathbb{E}[\phi(\bar{Z}_n^{(-I)}) \mid Z_1, \dots, Z_{n+1}].$$

Taking outer expectations yields the claim since $\mathbb{E}[\phi(\bar{Z}_n^{(-I)})|I = j] = \mathbb{E}[\phi(\bar{Z}_n)]$ for each j. □

5.1 Regular exponential families and Bregman form of reverse KL

Let ν be a measure on $\mathcal{X}$ and consider a (minimal, regular) exponential family on $\mathcal{X}$ given by the densities

$$p_\theta(x) = \exp\big(\langle \theta, T(x)\rangle - A(\theta)\big)\, h(x), \qquad \theta \in \Theta \subset \mathbb{R}^m,$$

where

$$A(\theta) = \log \int \exp(\langle \theta, T(x)\rangle) h(x)\, d\nu(x).$$

Define the mean map $\mu(\theta) := \mathbb{E}_\theta[T(X)] = \nabla A(\theta)$. Let $\mathcal{M} := \nabla A(\Theta)$ be the mean-parameter space, and let

$$A^*(\mu) := \sup_{\theta \in \Theta} \{\langle \theta, \mu\rangle - A(\theta)\}$$

be the Fenchel dual on $\mathcal{M}$, which is well known to be convex.

Proposition 5.2. *Let $\theta, \theta_* \in \Theta$, and set $\mu = \nabla A(\theta)$, $\mu_* = \nabla A(\theta_*)$. Then*

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta_*}) = A^*(\mu) - A^*(\mu_*) - \langle \nabla A^*(\mu_*),\, \mu - \mu_*\rangle =: B_{A^*}(\mu, \mu_*),$$

the Bregman divergence generated by the convex function A^*. In particular, $\mu \mapsto D_{\mathrm{KL}}(p_{\theta(\mu)} \| p_{\theta_*})$ is convex on $\mathcal{M}$.

Proof. We compute directly by expanding the log-density ratio. Since both p_θ and p_{θ_*} share the same base measure $h(x)\, d\nu(x)$, the h terms cancel in $\log(p_\theta / p_{\theta_*})$, leaving only the sufficient-statistic and log-partition contributions:

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta_*}) = \mathbb{E}_\theta[\langle \theta - \theta_*, T(X)\rangle - A(\theta) + A(\theta_*)] = \langle \theta - \theta_*, \mu\rangle - A(\theta) + A(\theta_*).$$

By Legendre duality, for $\mu = \nabla A(\theta)$ one has $A^*(\mu) = \langle \theta, \mu\rangle - A(\theta)$, and similarly $A^*(\mu_*) = \langle \theta_*, \mu_*\rangle - A(\theta_*)$. Moreover $\nabla A^*(\mu_*) = \theta_*$. Substituting these identities into the previous display and rearranging yields the stated Bregman form, which is clearly convex. □

Let $X_1, \ldots, X_n \sim p_{\theta_*}$ be IID and write $\bar{T}_n = \frac{1}{n}\sum_{i=1}^n T(X_i)$. Whenever the MLE exists and lies in the interior, it satisfies

$$\nabla A(\widehat{\theta}_n) = \bar{T}_n, \qquad \iff \qquad \widehat{\mu}_n := \nabla A(\widehat{\theta}_n) = \bar{T}_n.$$

We now prove Proposition 1.5.

Proof of Proposition 1.5. By Proposition 5.2,

$$D_{\mathrm{KL}}(p_{\widehat{\theta}_n} \| p_{\theta_*}) = B_{A^*}(\widehat{\mu}_n, \mu_*) = B_{A^*}(\bar{T}_n, \mu_*).$$

Define $\phi(\mu) := B_{A^*}(\mu, \mu_*)$. As a function of μ, this is convex because it is the sum of the convex function $A^*(\mu)$ and an affine function of μ (the remaining Bregman terms depend on μ only linearly).

Now $\bar{T}_n = \frac{1}{n}\sum_{i=1}^n T(X_i)$ is the sample mean of the IID vectors $T(X_i)$. Apply Lemma 5.1 with $Z_i := T(X_i)$ and $\bar{Z}_n := \bar{T}_n$ to conclude $\mathbb{E}[\phi(\bar{T}_{n+1})] \leq \mathbb{E}[\phi(\bar{T}_n)]$, i.e. the expected reverse KL is nonincreasing. Strictness follows from the strict convexity/non-degeneracy conditions exactly as in Lemma 5.1. □

5.2 Explicit Reverse KL for Gaussians with Known Mean

Assume $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma_*)$ and the MLE is

$$\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

We derive explicit formulas for the reverse-KL learning curve

$$\widetilde{\mathcal{E}}_{n,d}^{(0)}(\Sigma_*) := \mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathcal{N}(0, \widehat{\Sigma}_n) \,\|\, \mathcal{N}(0, \Sigma_*)\big)\Big]$$

and again deduce monotonicity.

Proof of Proposition 1.6. As before we reduce to the identity-covariance case by coordinate change. Set $Y_i := \Sigma_*^{-1/2} X_i$, so $Y_i \sim \mathcal{N}(0, I_d)$, and define the empirical covariance in the new coordinates by

$$\widehat{S}_n := \Sigma_*^{-1/2} \widehat{\Sigma}_n \Sigma_*^{-1/2} = \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^\top.$$

Since KL divergence is invariant under the invertible change of variables $x \mapsto \Sigma_*^{-1/2} x$,

$$D_{\mathrm{KL}}\big(\mathcal{N}(0, \widehat{\Sigma}_n) \,\|\, \mathcal{N}(0, \Sigma_*)\big) = D_{\mathrm{KL}}\big(\mathcal{N}(0, \widehat{S}_n) \,\|\, \mathcal{N}(0, I_d)\big).$$

The Gaussian KL formula gives, for $\Sigma \succ 0$, that

$$D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma) \| \mathcal{N}(0, I_d)) = \frac{1}{2}(\mathrm{tr}(\Sigma) - d - \log\det \Sigma)$$

and hence

$$D_{\mathrm{KL}}\big(\mathcal{N}(0, \widehat{S}_n) \,\|\, \mathcal{N}(0, I_d)\big) = \frac{1}{2}\Big(\, \mathrm{tr}(\widehat{S}_n) - d - \log\det(\widehat{S}_n)\Big).$$

Let $W := \sum_{i=1}^{n} Y_i Y_i^\top$, so $W \sim \mathrm{Wishart}_d(n, I_d)$ and $\widehat{S}_n = W/n$. Moreover $\mathrm{tr}(W) = \sum_{i=1}^{n} \|Y_i\|^2$, so $\mathbb{E}[\mathrm{tr}(W)] = nd$ and thus $\mathbb{E}[\mathrm{tr}(\widehat{S}_n)] = d$. Therefore the $\mathrm{tr}(\widehat{S}_n) - d$ term vanishes in expectation, giving

$$\widetilde{\mathcal{E}}_{n,d}^{(0)}(\Sigma_*) = -\frac{1}{2}\, \mathbb{E}[\log\det(\widehat{S}_n)] = -\frac{1}{2}\Big(\mathbb{E}[\log\det W] - d \log n\Big).$$

Applying Proposition 3.4 yields

$$\mathbb{E}[\log\det W] = d \log 2 + \sum_{j=1}^{d} \psi\Big(\frac{n-j+1}{2}\Big),$$

so

$$\widetilde{\mathcal{E}}_{n,d}^{(0)}(\Sigma_*) = \frac{1}{2}\, r_d(n), \qquad r_d(x) := d \log(x/2) - \sum_{j=1}^{d} \psi\Big(\frac{x-j+1}{2}\Big), \quad x > d-1. \tag{20}$$

If $n < d$, then W is singular a.s., so $\log\det(\widehat{S}_n) = -\infty$ and the expectation is $+\infty$.

For monotonicity, differentiate $r_d(x)$ for $x > d-1$:

$$r_d'(x) = \frac{d}{x} - \frac{1}{2}\sum_{j=1}^{d} \psi_1\Big(\frac{x-j+1}{2}\Big).$$

By (8), for each j,

$$\frac{1}{2}\psi_1\Big(\frac{x-j+1}{2}\Big) > \frac{1}{x-j+1}.$$

Summing over j gives

$$r_d'(x) < \frac{d}{x} - \sum_{j=1}^{d} \frac{1}{x-j+1} \le \frac{d}{x} - \frac{d}{x} = 0,$$

since each denominator satisfies $x-j+1 \le x$ and hence $(x-j+1)^{-1} \ge x^{-1}$. Hence r_d is strictly decreasing on $(d-1,\infty)$, and in particular $r_d(n+1) < r_d(n)$ for all integers $n \ge d$. □

5.3 Explicit Reverse KL for Gaussians with Unknown Mean

Assume $X_1,\dots,X_n \sim \mathcal{N}(\mu_*,\Sigma_*)$ and use the Gaussian MLEs

$$\widehat{\mu}_n := \bar{X}, \qquad \widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^\top.$$

Define the reverse-KL learning curve

$$\widetilde{\mathcal{E}}^{\text{full}}_{n,d}(\mu_*,\Sigma_*) := \mathbb{E}\Big[D_{\text{KL}}\big(\mathcal{N}(\widehat{\mu}_n,\widehat{\Sigma}_n)\,\|\,\mathcal{N}(\mu_*,\Sigma_*)\big)\Big].$$

Proof of Proposition 1.7. Again we use the coordinate change

$$Y_i := \Sigma_*^{-1/2}(X_i-\mu_*) \sim \mathcal{N}(0,I_d), \qquad \bar{Y} = \frac{1}{n}\sum_i Y_i.$$

Then

$$\widehat{\mu}_n - \mu_* = \Sigma_*^{1/2}\bar{Y}, \qquad \widehat{\Sigma}_n = \Sigma_*^{1/2}\widehat{S}_n\Sigma_*^{1/2}, \qquad \widehat{S}_n := \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})^\top.$$

The Gaussian KL formula gives

$$\begin{aligned} D_{\text{KL}}\big(\mathcal{N}(\widehat{\mu}_n,\widehat{\Sigma}_n)\,\|\,\mathcal{N}(\mu_*,\Sigma_*)\big) &= \frac{1}{2}\Big(\operatorname{tr}(\Sigma_*^{-1}\widehat{\Sigma}_n) + (\widehat{\mu}_n-\mu_*)^\top\Sigma_*^{-1}(\widehat{\mu}_n-\mu_*) - d - \log\det(\Sigma_*^{-1}\widehat{\Sigma}_n)\Big) \\ &= \frac{1}{2}\Big(\operatorname{tr}(\widehat{S}_n) + \|\bar{Y}\|^2 - d - \log\det(\widehat{S}_n)\Big), \end{aligned}$$

which is again independent of (μ_*,Σ_*).

Let

$$W := \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^\top$$

so $\widehat{S}_n = W/n$. By Lemma 3.7 (applied to $Y_1, \ldots, Y_n \sim \mathcal{N}(0, I_d)$), we have

$$W \sim \mathrm{Wishart}_d(n-1, I_d)$$

and that W is independent of $\bar{Y} \sim \mathcal{N}(0, I_d/n)$. Consequently,

$$\begin{aligned} \mathbb{E}[\mathrm{tr}(\widehat{S}_n)] &= \frac{1}{n}\mathbb{E}[\mathrm{tr}(W)] = \frac{d(n-1)}{n} = d\Big(1 - \frac{1}{n}\Big); \\ \mathbb{E}[\|\bar{Y}\|^2] &= \mathrm{tr}(\mathbb{E}[\bar{Y}\bar{Y}^\top]) = \mathrm{tr}(I_d/n) = \frac{d}{n}. \end{aligned}$$

Thus $\mathbb{E}[\mathrm{tr}(\widehat{S}_n) + \|\bar{Y}\|^2 - d] = 0$. In particular, after taking expectations the reverse KL reduces again to a (negative) log-determinant term independent of (μ_*, Σ_*):

$$\widetilde{\mathcal{E}}_{n,d}^{\mathrm{full}}(\mu_*, \Sigma_*) = -\frac{1}{2}\,\mathbb{E}[\log\det(\widehat{S}_n)] = -\frac{1}{2}\Big(\mathbb{E}[\log\det W] - d\log n\Big).$$

Applying Proposition 3.4 with n replaced by $n-1$ yields:

$$\mathbb{E}[\log\det W] = d\log 2 + \sum_{j=1}^{d} \psi\Big(\frac{(n-1)-j+1}{2}\Big) = d\log 2 + \sum_{j=1}^{d}\psi\Big(\frac{n-j}{2}\Big),$$

i.e. for r_d in (20) we have $\widetilde{\mathcal{E}}_{n,d}^{\mathrm{full}} = r_d(n-1)/2$. We have already seen that r_d is strictly decreasing on $(d-1, \infty)$; thus $\widetilde{\mathcal{E}}_{n,d}^{\mathrm{full}}$ is decreasing in n for $n > d$, concluding the proof. □

Acknowledgement Thanks to Sébastien Bubeck and Mehtaab Sawhney for helpful comments.

References

[AFS92] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation*, 4(4):605–618, 1992.

[Ama93] S. Amari. A universal theorem on learning curves. *Neural networks*, 6(2):161–166, 1993.

[And03] T. Anderson. An Introduction to Multivariate Statistical Analysis. *Wiley series in probability and statistics*, 2003.

[BDK+22] O. J. Bousquet, A. Daniely, H. Kaplan, Y. Mansour, S. Moran, and U. Stemmer. Monotone learning. In *Conference on Learning Theory*, pages 842–866. PMLR, 2022.

[BE53] H. Bateman and A. Erdélyi. Higher transcendental functions, volume II. *Bateman Manuscript Project) McGraw-Hill Book Company*, 410, 1953.

[Ber29] S. Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52(1):1–66, 1929.

[BHMM19] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, 116(32):15849–15854, 2019.

[BHX20] M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[CT99] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1999.

[DGL97] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1997.

[HBM$^+$22] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. In *36th International Conference on Neural Information Processing Systems*, 2022.

[HKOS91] D. Haussler, M. Kearns, M. Opper, and R. Schapire. Estimating average-case learning curves using Bayesian, statistical physics and VC dimension methods. *Advances in Neural Information Processing Systems*, 4, 1991.

[HMRT22] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949, 2022.

[KBJ19] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2019.

[KMH$^+$20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[Lax14] P. D. Lax. *Functional analysis.* John Wiley & Sons, 2014.

[LKB23] M. Loog, J. H. Krijthe, and M. Bicego. Also for k-means: more data does not imply better performance. *Machine Learning*, 112(8):3033–3050, 2023.

[LVM19] M. Loog, T. Viering, and A. Mey. Minimizers of the empirical risk and risk monotonicity. *Advances in Neural Information Processing Systems*, 32, 2019.

[LW21] Y. Li and Y. Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv:2110.09502*, 2021.

[Mer14] M. Merkle. Completely monotone functions: a digest. In *Analytic Number Theory, Approximation Theory, and Special Functions: In Honor of Hari M. Srivastava*, pages 347–364. 2014.

[MM22] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022.

[Mui09] R. J. Muirhead. *Aspects of multivariate statistical theory.* John Wiley & Sons, 2009.

[NKB$^+$21] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[OH91] M. Opper and D. Haussler. Calculation of the learning curve of bayes optimal classification algorithm for learning a perceptron with noise. In *COLT*, volume 91, pages 75–87, 1991.

[PC87] S. Patarnello and P. Carnevali. Learning networks of neurons with boolean logic. *Europhysics Letters*, 4(4):503, 1987.

[Pes22] V. Pestov. A universally consistent learning rule with a universally monotone error. *Journal of Machine Learning Research*, 23(157):1–27, 2022.

[SST92] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[TT93] H. Takahashi and E. Tomita. Estimating learning curves by PAC-learnability criterion. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1641–1644. IEEE, 1993.

[VL22] T. Viering and M. Loog. The shape of learning curves: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2022.

[VML19] T. Viering, A. Mey, and M. Loog. Open problem: Monotonicity of learning. In *Conference on Learning Theory*, pages 3198–3201. PMLR, 2019.

[YHB$^+$21] G. Yang, E. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.