Gradient conjugate priors and deep neural networks

Pavel Gurevich*, Hannes Stuke†

February 9, 2018

Abstract

The paper deals with learning the probability distribution of the observed data by artificial neural networks. We suggest a so-called gradient conjugate prior (GCP) update appropriate for neural networks, which is a modification of the classical Bayesian update for conjugate priors. We establish a connection between the gradient conjugate prior update and the maximization of the log-likelihood of the predictive distribution. Unlike for the Bayesian neural networks, we do not impose a prior on the weights of the neural networks, but rather assume that the ground truth distribution is normal with unknown mean and variance and learn by neural networks the parameters of a prior (normal-gamma distribution) for these unknown mean and variance. The update of the parameters is done, using the gradient that, at each step, directs towards minimizing the Kullback–Leibler divergence from the prior to the posterior distribution (both being normal-gamma). We obtain a corresponding dynamical system for the prior's parameters and analyze its properties. In particular, we study the limiting behavior of all the prior's parameters and show how it differs from the case of the classical full Bayesian update. The results are validated on synthetic and real world data sets.

Keywords. Conjugate priors, Kullback–Leibler divergence, Student's t-distribution, deep neural networks, regression, uncertainty quantification, asymptotics, outliers

Contents

*Free University of Berlin, Arnimallee 3, 14195 Berlin, Germany; RUDN University, Miklukho-Maklaya 6, 117198 Moscow, Russia; email: gurevich@math.fu-berlin.de

†Free University of Berlin, Arnimallee 7, 14195 Berlin, Germany; email: h.stuke@fu-berlin.de

1 Introduction

Reconstructing the probability distribution of the data by artificial neural networks based upon observations is one of the main approaches to quantification of uncertainty of predictions. Under the assumption that the data are normally distributed, the most well studied way of reconstructing the probability distribution is the Bayesian learning of neural networks [18]. One treats the weights of the network as normally distributed random variables, imposes their prior distribution, and then finds the posterior distribution conditioned on the data. The main difficulty is that neither the posterior, nor the resulting predictive distributions are given in a closed form. As a result, different approximation methods have been developed [3, 7, 8, 10, 22]. However, many of them have certain drawbacks related to the lack of scalability in data size or the neural network complexity, and are still a field of

ongoing research, see, the recent paper [7]. Among other methods for uncertainty quantification, there are the *delta method* [9, 30, 31], the *mean-variance estimate* [23], and the *deep ensembles* [14, 15]. A combination of the Bayesian approach (using the dropout variational inference) with the mean-variance estimate was used in [12]. A new method based on minimizing a joint loss for a regression network and another network quantifying uncertainty was recently proposed in [5]. We refer to [13, 27] and the recent works [3, 5, 14, 15, 20] for a comprehensive comparison of the above methods and further references to research on the Bayesian learning of neural networks.

We study an alternative approach to reconstructing the ground truth probability distribution based on what we call a *gradient conjugate prior (GCP) update*. Assuming that the ground truth distribution $q(y)$ of the data $\mathbf{y}$ is normal with unknown mean and precision, we let neural networks (with deterministic weights) learn the four parameters of the normal-gamma distribution that serves as a prior for the mean and variance of $\mathbf{y}$. Given a new observation, the classical Bayesian update yields the posterior distribution for the mean and variance of $\mathbf{y}$. This posterior appears normal-gamma as well [2]. However, *one cannot update its parameters directly* because they are represented by the outputs of the neural networks. Instead, one has to update the weights of the neural networks. We suggest to make a gradient descent step in the direction of minimization of the Kullback–Leibler (KL) divergence from the prior to the posterior (see the details in Sec. 2.4). This is the step that we call the GCP update. After updating the weights, one takes the next observation and repeats the above update procedure. At each step, one has the predictive distribution in the form of a (non-standardized) Student's t-distribution $p_{\text{pred}}(y)$, whose parameters are explicitly determined by the outputs of the neural networks.

In the paper, we provide a detailed analysis of the dynamics given by the GCP update. Intuitively, one might think that the GCP update, after convergence, yields the same result as the classical CP update. Surprisingly, this is not quite the case (see Remark 3.4). The first observation, which we prove in Sec. 2.5, is that the *GCP update is actually equivalent to maximizing by gradient ascent the likelihood of the predictive distribution* $p_{\text{pred}}(y)$. As the number of observations tends to infinity the *GCP update becomes also equivalent to minimizing by gradient descent the KL divergence from the predictive distribution* $p_{\text{pred}}(y)$ *to the ground truth distribution* $q(y)$. We show that these equivalences hold in general, even if the prior is not conjugate to the likelihood function.

Now let us come back to our original assumption that $q(y)$ is a normal distribution and $p_{\text{pred}}(y)$ is a Student's t-distribution. The latter appears to be overparametrized (by four parameters instead of three). We keep it overparametrized in order to compare the dynamics of the parameters under the classical CP update and under the GCP update. Reformulation of our results for Student's t-distribution parameterized in the standard way by three parameters will be straightforward. There is a vast literature on the estimation of parameters of Student's t-distribution, see, e.g., the overview [21] and the references therein. Note that, in the context of neural networks, different samples correspond to different inputs of the network, and hence they belong to different Student's t-distributions with *different unknown parameters*. Thus, the maximization of the likelihood of Student's t-distribution

with respect to the weights of the networks is one of the most common methods. In [29], the possibility of utilizing evolutionary algorithms for maximaizing the likelihood was explored experimentally. Another natural way is to use the gradient ascent with respect to the weights of the network. As we said, the latter is equivalent to the usage of the GCP update. In the paper, we obtain a dynamical system for the prior's parameters that approximates the GCP update (as well as the gradient ascent for maximization of Student's t-distribution). We study the dynamics of the prior's parameters in detail, in particular analyzing their convergence properties. Our approach is illustrated with synthetic data and validated on various real-world data sets in comparison with other methods for learning probability distributions based on neural networks. To our best knowledge, neither the dynamical systems analysis of the GCP (or gradient ascent for maximizing the likelihood of Student's t-distribution), nor a thorough comparison of the GCP with other methods has been carried out before.

As an interesting and useful consequence of our analysis, we will see how the GCP interacts with outliers (a small percentage of observations that do not come from the assumed normal distribution $q(y)$). The outliers prevent one of the prior's parameters (α, which is related to the number of degrees of freedom of $p_{\text{pred}}(y)$) from going to infinity. On one hand, this is known [17, 25] to allow for a better estimate of the mean and variance of $q(y)$, compared with directly using the maximization of the likelihood of a normal distribution. On the other hand, this still leads to overestimation of the variance of $q(y)$. To deal with this issue, we obtain an explicit formula that allows one to correct the estimate of the variance and recover the ground truth variance of $q(y)$. To our knowledge, such a correction formula was not derived in the literature before.

The paper is organized as follows. In Sec. 2, we provide a detailed motivation for the GCP update, explain how we approximate the parameters of the prior distribution by neural networks, establish the relation between the GCP update and the predictive distribution, and formulate the method of learning the ground truth distribution from the practical point of view. Section 3 is the mathematical core of this paper. We derive a dynamical system for the prior's parameters, induced by the GCP update, and analyze it in detail. In particular, we obtain an asymptotics for the growth rate of α and find the limits of the other parameters of the prior. In Sec. 4, we study the dynamics for a fixed α. We find the limiting values for the rest of the parameters and show how one can recover the variance of the ground truth normal distribution $q(y)$. In Sec. 5, we compare the sensitivity to outliers in the GCP update with that in minimizing the standard squared error loss or maximizing the log-likelihood of a normal distribution. Finally, we show how α controls the learning speed in clean and noisy regions. In Sec. 6, we illustrate the fit of neural networks for synthetic and various real-world data sets. Section 7 contains a conclusion and an outline of possible directions of further research. Appendix A contains auxiliary technical results that are used for finding asymptotics in Sec. 3 and the variance of the ground truth distribution in Sec. 4.

2 Motivation

2.1 Estimating normal distributions with unknown mean and precision

Assume one wants to estimate unknown mean $\mu \in \mathbb{R}$ and unknown precision $\tau > 0$ (the inverse of the variance) of normally distributed data $\mathbf{y}$. One standard approach for estimating μ and τ is based on the conjugate prior update. One assumes that μ and τ have a joint prior given by the normal-gamma distribution

$$p(\mu, \tau; m, \nu, \alpha, \beta) = \frac{\beta^\alpha \nu^{1/2}}{\Gamma(\alpha)(2\pi)^{1/2}} \tau^{\alpha - 1/2} e^{-\beta\tau} e^{-\frac{\nu\tau(\mu-m)^2}{2}}, \tag{2.1}$$

where $m \in \mathbb{R},\ \nu > 0,\ \alpha > 1,\ \beta > 0$.

The marginal distribution for μ is a non-standardized Student's t-distribution, and we have

$$\mathbb{E}[\mu] = m, \quad \mathbb{V}[\mu] = \frac{\beta}{\nu(\alpha - 1)} \tag{2.2}$$

The marginal distribution for τ is the Gamma distribution, and we have

$$\mathbb{E}[\tau] = \frac{\alpha}{\beta}, \quad \mathbb{V}[\tau] = \frac{\mathbb{E}(\tau)}{\beta} = \frac{\alpha}{\beta^2}. \tag{2.3}$$

By marginalizing τ and μ, one can get the predictive distribution $p_{\text{pred}}(y) = p_{\text{pred}}(y; m, \nu, \alpha, \beta)$ for $\mathbf{y}$, which appears to be a non-standardized Student's t-distribution. Its mean and variance can be used to estimate the mean and variance of $\mathbf{y}$. The estimated mean m_{est} and variance V_{est} are given by

$$m_{\text{est}} := m, \quad V_{\text{est}} := \frac{\beta(\nu + 1)}{(\alpha - 1)\nu}. \tag{2.4}$$

We refer, e.g., to [2] for further details.

2.2 Conjugate prior updade

Suppose one observes a new sample $\overline{y}$. Then, by the Bayes theorem, the conditional distribution of (μ, τ) under the condition that $\mathbf{y} = \overline{y}$ (called posterior distribution and denoted by $p_{\text{post}}(\mu, \tau)$) appears to be normal-gamma as well [2], namely,

$$p_{\text{post}}(\mu, \tau) = p(\mu, \tau; m', \nu', \alpha', \beta', \overline{y}), \tag{2.5}$$

where the parameters are updated as follows:

$$m' = \frac{\nu m + \overline{y}}{\nu + 1}, \quad \nu' = \nu + 1, \quad \alpha' = \alpha + \frac{1}{2}, \quad \beta' = \beta + \frac{\nu}{\nu + 1} \frac{(\overline{y} - m)^2}{2}. \tag{2.6}$$

We call (2.6) the *conjugate prior (CP) update.*

2.3 Kullback–Leibler divergence

The Kullback–Leibler (KL) divergence from a continuous distribution p to a continuous distribution p_{post} is defined as follows:

$$D_{\text{KL}}(p_{\text{post}}\|p) := \int p_{\text{post}}(\mu,\tau)\,\ln\frac{p_{\text{post}}(\mu,\tau)}{p(\mu,\tau)}\,d\mu\,d\tau. \tag{2.7}$$

We denote by $\Psi(x) := \Gamma'(x)/\Gamma(x)$ the digamma function, where $\Gamma(x)$ is the gamma function. Then for the above normal-gamma distributions (2.1) and (2.5) the KL divergence takes the form [26]

$$\begin{aligned} K(m,\nu,\alpha,\beta) := {} & \frac{1}{2}\frac{\alpha'}{\beta'}(m-m')^2\nu + \frac{1}{2}\frac{\nu}{\nu'} - \frac{1}{2}\ln\frac{\nu}{\nu'} - \frac{1}{2} \\ & - \alpha\ln\frac{\beta}{\beta'} + \ln\frac{\Gamma(\alpha)}{\Gamma(\alpha')} - (\alpha-\alpha')\Psi(\alpha') + (\beta-\beta')\frac{\alpha'}{\beta'}. \end{aligned} \tag{2.8}$$

2.4 Approximation of the parameters by gradient conjugate prior neural networks

Our goal is to approximate the parameters m,α,β,ν by deep neural networks, i.e., to represent them as functions of weights: $m = m(w_1), \alpha = \alpha(w_2), \beta = \beta(w_3), \nu = \nu(w_4)$, $w_j \in \mathbb{R}^{M_j}$. (They are also functions of an input variable x, but we do not indicate this explicitly.) In this case, one cannot directly apply the update in (2.6), but has to update the weights w_j instead. The natural way to do so is to observe a sample $\overline{y}$, to calculate the posterior distribution (2.5) and to change the weights w in the direction of $-\nabla_w K$, i.e.,

$$w_{\text{new}} := w - \lambda\cdot\nabla_w K, \quad w \in \{w_1,\dots,w_4\}, \tag{2.9}$$

where $\lambda > 0$ is a fixed learning rate. When we compute the gradient of K with respect of w, we keep all the prime variables in (2.8) fixed and do not treat them as functions of w, while all the nonprime variables are treated as functions of w. We still use the notation $\nabla_w K$ in this case. We call (2.9) the *gradient conjugate prior (GCP) update*.

As we will see below, this update induces the update for (m,α,β,ν) that is different from the classical conjugate prior update (2.6) and yields a completely different dynamics. Before we analyze this dynamics in detail, we present an alternative viewpoint on the GCP update, which provides an insight into what is actually optimized by (2.9) in the general case.

2.5 GCP update and learning the predictive distribution

Suppose we want to learn a ground truth probability distribution $q(y)$ of a random variable $\mathbf{y}$ (a normal distribution in our particular case). Since the ground truth distribution is a priori unknown, we conjecture that it belongs to a family of distributions $L(y;\boldsymbol{\tau})$ parametrized by $\boldsymbol{\tau}$ (in our case $\boldsymbol{\tau} = (\mu,\tau)$ and $L(y;\boldsymbol{\tau})$ is a normal distribution with mean μ and precision τ). Since $\boldsymbol{\tau}$ is a priori unknown, we assume it is a random variable with the prior distribution from

the family $p(\boldsymbol{\tau}; w)$ parametrized by w (in our case, $p(\boldsymbol{\tau}; w)$ is the normal-gamma distribution and w are the weights of the neural networks approximating m, α, β, ν, see Sec. 2.4). We denote the predictive distribution by

$$p_{\text{pred}}(y; w) := \int L(y, \boldsymbol{\tau}) p(\boldsymbol{\tau}; w)\, d\boldsymbol{\tau} \tag{2.10}$$

(non-standardized Student's t-distribution in our case). Given an observation $\mathbf{y} = y$, the Bayes rule determines the posterior distribution

$$p_{\text{post}}(\boldsymbol{\tau}; w, y) := \frac{L(y; \boldsymbol{\tau}) p(\boldsymbol{\tau}; w)}{p_{\text{pred}}(y; w)}. \tag{2.11}$$

In our case, $p_{\text{post}}(\boldsymbol{\tau}; w, \overline{y})$ is normal-gamma again, but we emphasize that, in general, it need not be from the same family as the prior $p(\boldsymbol{\tau}; w)$ is.

Now we compute the gradient of the KL divergence

$$K(w, y) := D_{\text{KL}}(p_{\text{post}} \| p) = \int p_{\text{post}}(\boldsymbol{\tau}; w, y) \ln \frac{p_{\text{post}}(\boldsymbol{\tau}; w, y)}{p(\boldsymbol{\tau}; w)}\, d\boldsymbol{\tau} \tag{2.12}$$

(cf. (2.7)) with respect to w, assuming that w in the posterior distribution is *freezed*, and we do not differentiate it. Denoting such a gradient by $\nabla_w K(w, y)$, we obtain the following lemma.

Lemma 2.1. $\nabla_w K(w, y) = -\nabla_w \ln p_{\text{pred}}(y; w)$.

Proof. Freezing $p_{\text{post}}(\boldsymbol{\tau}; w, y)$ in (2.12), we have

$$\nabla_w K(w, y) = -\int \frac{p_{\text{post}}(\boldsymbol{\tau}; w, y) \nabla_w p(\boldsymbol{\tau}; w)}{p(\boldsymbol{\tau}; w)}\, d\boldsymbol{\tau}.$$

Plugging in $p_{\text{post}}(\boldsymbol{\tau}; w, y)$ from (2.11) and using (2.10) yields

$$\nabla_w K(w, y) = -\int \frac{L(y, \boldsymbol{\tau}) \nabla_w p(\boldsymbol{\tau}; w)}{p_{\text{pred}}(y; w)}\, d\boldsymbol{\tau} = -\frac{\nabla_w p_{\text{pred}}(y; w)}{p_{\text{pred}}(y; w)} = -\nabla_w \ln p_{\text{pred}}(y; w).$$

□

Lemma 2.1 shows that the GCP update (2.9) is the gradient ascent step in the direction of maximizing the log-likelihood of the predictive distribution $p_{\text{pred}}(y; w)$ given a new observation $\mathbf{y} = y$. Furthermore, using Lemma 2.1, we see that given observations $y_1, \ldots, y_N$, the averaged GCP update of the parameters w is given by (cf. (2.9))

$$w_{\text{new}} := w - \lambda \frac{1}{N} \sum_{n=1}^{N} [\nabla_w K(w, y_n)] = w + \lambda \nabla_w \left(\frac{1}{N} \sum_{n=1}^{N} \ln p_{\text{pred}}(y_n; w) \right). \tag{2.13}$$

Further, if the observations are sampled from the ground truth distribution $q(y)$ and its number tends to infinity, then the GCP update (2.13) assumes the form

$$\begin{aligned} w_{\text{new}} &:= w - \lambda \cdot \mathbb{E}_{y\sim q(y)}[\nabla_w K(w,y)] = w - \lambda \nabla_w \int \ln p_{\text{pred}}(y;w) q(y)\, dy \\ &= w - \lambda \nabla_w \int q(y) \ln \frac{q(y)}{p_{\text{pred}}(y;w)}\, dy = w - \lambda \nabla_w D_{\text{KL}}(q \| p_{\text{pred}}(\cdot;w)). \end{aligned} \tag{2.14}$$

Remark 2.1. 1. Formula (2.13) shows that the GCP update *maximizes the likelihood of the predictive distribution* $p_{\text{pred}}(y;w)$ *for the the observations* $y_1, \dots, y_N$.

2. Formula (2.14) shows that the GCP update is equivalent to the gradient descent step for the *minimization of the KL divergence from the predictive distribution* $p_{\text{pred}}(y;w)$ *to the ground truth distribution* $q(y)$. If the ground truth distribution $q(y)$ belongs to the family $p_{\text{pred}}(y;w)$, then the minimum equals zero and is achieved for some (not necessarily unique) w_* such that $p_{\text{pred}}(y;w_*) = q(y)$; otherwise the minimum is positive and provides the best possible approximation of the ground truth in the sense of the KL divergence.

3. In our case, $q(y)$ is a normal distribution and $p_{\text{pred}}(y;w)$ are Student's t-distributions. In accordance with item 2, we will see below that the GCP update forces the number of degrees of freedom of $p_{\text{pred}}(y;w)$ tend to infinity. However, due to the over-parametrization of the predictive distribution (four parameters m, α, β, ν instead of three), the learned variance of $q(y)$ will be represented by a curve in the space (β, ν). The limit point β_*, ν_* to which β, ν will converge during the GCP update, will depend on the initial condition. Interestingly, β_*, ν_* will always be different from the limit point obtained by the classical CP update (2.6) (cf. Remark 3.4).

2.6 Practical approaches

Based on Remark 2.1 (items 1 and 2), we suggest the following general practical approach.

Practical approach 2.1. *1. One approximates the parameters of the prior by neural networks:*

$$m = m(w_1),\ \alpha = \alpha(w_2),\ \beta = \beta(w_3),\ \nu = \nu(w_4). \tag{2.15}$$

We call them the GCP neural networks.

2. One trains these four networks by the GCP update (2.9) *until convergence of* m, α, β, ν.

3. The resulting predictive distribution is the non-standardized Student's t-distribution $t_{2\alpha}(y|m, \beta(\nu+1)/(\nu\alpha))$. *The estimated mean* m_{est} *and variance* V_{est} *(for* $\alpha > 1$*) are given by*

$$m_{\text{est}} := m, \quad V_{\text{est}} := \frac{\beta(\nu+1)}{(\alpha-1)\nu}. \tag{2.16}$$

In practice one has finitely many observations y_n, $n = 1, \dots, N$, and the distribution $q(y)$ is a linear combination of the Dirac delta functions supported at y_n. Due to Remark 2.1 (item 1), Approach 2.1 yields the predictive Student's t-distribution with maximal likelihood. However, if the observations are sampled from a normal distribution and their number tends to infinity, α will tend to infinity due to Remark 2.1 (item 3). We will also show that $\nu \to 0$, and β will converge to a finite value $\beta_* > 0$.

Remark 2.2. Below we will justify the fact that if α is fixed and equals to some value α_*, then we obtain the best approximation of $q(y)$ by a non-standardized Student's t-distribution with $2\alpha_*$ degrees of freedom. However, one can still recover the correct variance of the ground truth normal distribution $q(y)$ by appropriately modifying the predictive variance V_{est} in (2.16), namely, by using

$$\tilde{V}_{\text{est}} := \frac{\beta_*(\nu_* + 1)}{(\alpha_* - A(\alpha_*))\nu_*} \tag{2.17}$$

with $A(\alpha_*)$ from Definition 3.1.

Furthermore, we will see that if the data in the training set come from a normal distribution but contain a small number of outliers in a certain region, then the GCP will automatically learn *finite values of* α in this region. This will lead to V_{est} that is higher than the ground truth variance of the normal distribution. If one does not expect the outliers in the test set, the variance estimate can be corrected by using (2.17) instead. This situation is illustrated in section 5.1, Fig. 5.1 and section 6.2, Fig. 6.1.

In the rest of the paper, we will rigorously justify the above approach based on the GCP update, study the dynamics of m, α, β, ν under this update, and analyze how one should correct the variance for a fixed α.

3 Dynamics of m, α, β, ν

3.1 Dynamical system for m, α, β, ν

The GCP update (2.9) induces the update for (m, α, β, ν) as follows:

$$\beta_{\text{new}} := \beta(w - \lambda \cdot \nabla_w K) = \beta\left(w - \lambda \frac{\partial K}{\partial \beta} \nabla_w \beta(w)\right) \approx \beta(w) - \lambda (\nabla_w \beta)^T (\nabla_w \beta) \frac{\partial K}{\partial \beta}, \tag{3.1}$$

where $w = w_3$, and similarly for m, α, ν.

Obviously, the new parameters $m_{\text{new}}, \alpha_{\text{new}}, \beta_{\text{new}}, \nu_{\text{new}}$ are different from $m', \alpha', \beta', \nu'$ given by the classical conjugate prior update (2.6). From now on, we replace $\lambda(\nabla_w \beta)^T(\nabla_w \beta)$ by a new learning rate and analyze *how the parameters will change and to which values they will converge under the updates of the form*

$$m_{\text{new}} := m - \lambda_1 \frac{\partial K}{\partial m}, \quad \alpha_{\text{new}} := \alpha - \lambda_2 \frac{\partial K}{\partial \alpha}, \quad \beta_{\text{new}} := \beta - \lambda_3 \frac{\partial K}{\partial m}, \quad \nu_{\text{new}} := \nu - \lambda_4 \frac{\partial K}{\partial \nu}, \tag{3.2}$$

where $\lambda_j > 0$ are the fixed learning rates. As before, when we compute the derivatives of K, we keep all the prime-variables in (2.8) fixed and do not treat them as functions of m, ν, α, β. In other words, *we first compute the derivatives of K with respect to m, ν, α, β and then substitute $m', \nu', \alpha', \beta'$ from* (2.6). For brevity, we will simply write $\partial K/\partial m$, etc. We call (3.2) the *GCP update* as well.

Setting

$$\sigma := \frac{\beta(\nu+1)}{\nu}, \tag{3.3}$$

we have

$$\frac{\partial K}{\partial m} = \frac{\alpha + 1/2}{\sigma + \frac{(m-\overline{y})^2}{2}}(m - \overline{y}), \tag{3.4}$$

$$\frac{\partial K}{\partial \alpha} = \ln\left(1 + \frac{(m-\overline{y})^2}{2\sigma}\right) + \Psi(\alpha) - \Psi\left(\alpha + \frac{1}{2}\right), \tag{3.5}$$

$$\frac{\partial K}{\partial \beta} = \frac{1}{\beta}\left(\frac{\alpha + 1/2}{1 + \frac{(m-\overline{y})^2}{2\sigma}} - \alpha\right), \tag{3.6}$$

$$\frac{\partial K}{\partial \nu} = \frac{1}{2\nu(\nu+1)}\left(\frac{\alpha + 1/2}{\sigma + \frac{(m-\overline{y})^2}{2}}(m - \overline{y})^2 - 1\right). \tag{3.7}$$

In this section and in the next one, we will treat the parameters μ, α, β, ν as functions of time $t > 0$ and study a dynamical system that approximates the GCP update (3.2) when the number of observations is large. We will concentrate on the prototype situation, where all new learning rates are the same.

Condition 3.1. *In the GCP update* (3.2)*, we have $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$.*

Under Condition 3.1, the approximating dynamical system takes the form

$$\dot{m} = -\mathbb{E}\left[\frac{\partial K}{\partial m}\right], \quad \dot{\alpha} = -\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right], \quad \dot{\beta} = -\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right], \quad \dot{\nu} = -\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]; \tag{3.8}$$

hereinafter the expectations are taken with respect to the true distribution $q(y)$ of $\mathbf{y}$ that is treated as a normally distributed random variable with mean $\mathbb{E}[\mathbf{y}]$ and variance $V := \mathbb{V}[\mathbf{y}]$, see Fig. 3.1.

Remark 3.1. Due to (2.14), system (3.8) defines the gradient flow with the potential $D_{\mathrm{KL}}(q \| p_{\mathrm{pred}}(\cdot; w))$, where $p_{\mathrm{pred}}(\cdot; w))$ is the probability distribution of the Student's t-distribution $t_{2\alpha}(y|m, \beta(\nu+1)/(\nu\alpha))$.

Remark 3.2. If Condition 3.1 does not hold, then the respective factors λ_j will appear in the right-hand sides in (3.8). The modifications one has to make in the arguments below are straightforward.

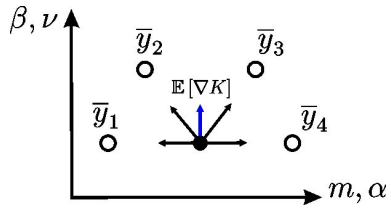

Figure 3.1: The black circle indicates the prior probability distribution (2.1) in the space of the parameters m, α, β, ν. The white circles indicate the posterior probability distributions (2.5) corresponding to different observations $\overline{y}_1, \overline{y}_2, \ldots$. The black vectors are the gradients with respect to the nonprime variables of the corresponding KL divergences. The blue vector is the averaged gradient. An equilibrium of system (3.8) would correspond to the case where the blue vector vanishes. Theorem 3.2 shows that this actually never happens. However, Theorem 4.1 shows that if one keeps α fixed, but updates m, β, ν, then one obtains a whole curve of equilibria.

3.2 Estimation of the mean m

Using (3.4), we obtain the formula for the expectation

$$\mathbb{E}\left[\frac{\partial K}{\partial m}\right] = \frac{\alpha + 1/2}{(2\pi V)^{1/2}} \int\limits_{-\infty}^{\infty} \frac{m - y}{\sigma + \frac{(m-y)^2}{2}} e^{-\frac{(\mathbb{E}[\mathbf{y}]-y)^2}{2V}} dy. \tag{3.9}$$

Theorem 3.1. *The first equation in* (3.8) *has a unique equilibrium* $m = \mathbb{E}[\mathbf{y}]$. *It is stable in the sense that, for any* α, β, ν, *we have*

$$\dot{m} < 0 \text{ if } m > \mathbb{E}[\mathbf{y}], \quad \dot{m} > 0 \text{ if } m > \mathbb{E}[\mathbf{y}].$$

Proof. Without loss of generality, assume that $\mathbb{E}[\mathbf{y}] = 0$ and $V = 1$ (otherwise, make a change of variables $z = (y - \mathbb{E}[\mathbf{y}])/V^{1/2}$ in the integral in (3.9)). Then we obtain from (3.9)

$$\mathbb{E}\left[\frac{\partial K}{\partial m}\right] = C_1 \int\limits_{-\infty}^{\infty} \frac{m - y}{C_2 + (m-y)^2} e^{-\frac{y^2}{2}} dy = -C_1 \int\limits_{-\infty}^{\infty} \frac{z}{C_2 + z^2} e^{-\frac{(m+z)^2}{2}} dz, \tag{3.10}$$

where $C_1, C_2 > 0$ do not depend on m.

Obviously, the right-hand side in (3.10) vanishes at $m = 0$. Furthermore, due to (3.10), for $m > 0$,

$$\mathbb{E}\left[\frac{\partial K}{\partial m}\right] = C_1 \int\limits_{0}^{\infty} \frac{z}{C_2 + z^2} \left(e^{-\frac{(m-z)^2}{2}} - e^{-\frac{(m+z)^2}{2}} \right) dz > 0$$

because $-(m-z)^2 > -(m+z)^2$ for $m, z > 0$. Similarly, $\mathbb{E}\left[\frac{\partial K}{\partial m}\right] < 0$ for $m < 0$. $\square$

3.3 Estimation of the variance. The unbounded absorbing set

From now on, taking into account Theorem 3.1, we assume the following.

Condition 3.2. $m = \mathbb{E}[\mathbf{y}]$.

Under Condition 3.2, we study the other three equations in (3.8), namely,

$$\dot{\alpha} = -\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right], \quad \dot{\beta} = -\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right], \quad \dot{\nu} = -\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right], \tag{3.11}$$

where (due to Condition 3.2)

$$\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right] = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \ln\left(1 + \frac{V}{\sigma}\frac{z^2}{2}\right) e^{-\frac{z^2}{2}} dz + \Psi(\alpha) - \Psi\left(\alpha + \frac{1}{2}\right), \tag{3.12}$$

$$\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right] = \frac{1}{\beta}\left(\frac{\alpha + 1/2}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{1}{1 + \frac{V}{\sigma}\frac{z^2}{2}} e^{-\frac{z^2}{2}} dz - \alpha\right), \tag{3.13}$$

$$\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right] = \frac{1}{2\nu(\nu+1)}\left(\frac{2\alpha+1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{z^2}{\frac{2\sigma}{V} + z^2} e^{-\frac{z^2}{2}} dz - 1\right). \tag{3.14}$$

3.3.1 The functions $A(\alpha)$ and $\sigma_{\varkappa}(\alpha)$

To formulate the main theorem of this section, we introduce a function $A(\alpha)$, which plays the central role throughout the paper.

Definition 3.1. For each $\alpha > 0$, $A(\alpha)$ is defined as a unique root of the equation

$$F(\alpha - A) = \frac{\alpha}{(2\alpha+1)(\alpha - A)}, \tag{3.15}$$

where

$$F(x) := \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{1}{2x + z^2} e^{-\frac{z^2}{2}} dz \ \left(= \frac{\pi^{1/2}}{2} \frac{e^x \mathrm{erfc}(x^{1/2})}{x^{1/2}}\right), \quad x > 0, \tag{3.16}$$

and erfc is the complementary error function.

The main properties of $A(\alpha)$ are given in the following lemma (see Fig. 3.2).

Lemma 3.1. *1. Equation* (3.16) *has a unique root* $A(\alpha)$*,*

2. $A(\alpha)$ *is monotonically increasing,*

3. $\frac{2\alpha}{2\alpha+3} < A(\alpha) < \min(\alpha, 1)$*,*

4. $A(\alpha)$ *satisfies the differential equation*

$$A'(\alpha) = 1 - \frac{2(\alpha - A)}{(2\alpha+1)A}, \tag{3.17}$$

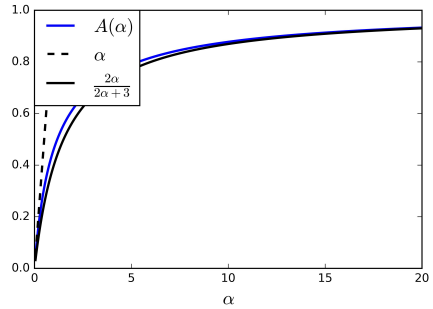

Figure 3.2: The function $A(\alpha)$ from Definition 3.1.

5. *$A(\alpha)$ has the following asymptotics:*

$$\begin{aligned} A(\alpha) &= \alpha - k_0\alpha^2 + o(\alpha^2) \ \text{ as } \alpha \to 0, \\ A(\alpha) &= 1 - \frac{k_1}{\alpha} + o\left(\frac{1}{\alpha}\right) \ \text{ as } \alpha \to \infty, \end{aligned} \tag{3.18}$$

where $k_0 = 4/\pi$, $k_1 = 3/2$.

Proof. These properties are proved in Lemmas A.1–A.4. □

Definition 3.2. For each $\varkappa \geq 0$, we define the functions (see Fig. 3.3, left)

$$\sigma_\varkappa(\alpha) := \left(1 - \frac{\varkappa}{\alpha}\right)(\alpha - A(\alpha))V, \quad \alpha > 0. \tag{3.19}$$

We remind that $V = \mathbb{V}[\mathbf{y}]$.

3.3.2 Estimation of the variance

The main result of this section (illustrated by Figures 3.3 and 3.4) is as follows.

Theorem 3.2. 1. *There is a smooth increasing function $\sigma_*(\alpha)$, $\alpha > 0$, such that*

(a) *$\dot{\alpha} = 0$ on the curve $(\alpha, \sigma_*(\alpha))$,*

(b) *$\lim_{\alpha\to 0} \sigma_*(\alpha) = 0$ and $\lim_{\alpha\to\infty} \sigma_*(\alpha) = \infty$,*

(c) *$\sigma_*(\alpha) < \sigma_0(\alpha)$ for all $\alpha > 0$,*

(d) *for any $\varkappa > 0$, there exists $\alpha_\varkappa > 0$ such that*

$$\sigma_*(\alpha) > \sigma_\varkappa(\alpha) \quad \text{for all } \alpha > \alpha_\varkappa,$$

(e) *the region*

$$\mathbf{S}_* := \left\{(\alpha, \beta, \nu) \in \mathbb{R}^3 : \alpha, \beta, \nu > 0 \text{ and } \sigma_*(\alpha) < \frac{\beta(\nu+1)}{\nu} < \sigma_0(\alpha)\right\} \tag{3.20}$$

is forward invariant for system (3.11).

2. *For any $\alpha(0), \beta(0), \nu(0) > 0$, there exists a time moment t_0 depending on the initial condition such that for all $t > t_0$, $(\alpha(t), \beta(t), \nu(t)) \in \mathbf{S}_*$, $\dot{\alpha}(t), \dot{\beta}(t) > 0$, $\dot{\nu}(t) < 0$.*

3. *For any $\alpha(0), \beta(0), \nu(0) > 0$, there is $C > 0$ depending on the initial conditions such that the points $(\nu(t), \beta(t))$ for all $t \geq 0$ lie on the integral curve*

$$\beta^2 + \nu^2 + \frac{2\nu^3}{3} = C \tag{3.21}$$

of the equation

$$\frac{d\beta}{d\nu} = -\frac{\nu(\nu+1)}{\beta}. \tag{3.22}$$

4. *For any $\alpha(0), \beta(0), \nu(0) > 0$, we have*

$$\alpha(t) \to \infty, \ A(\alpha(t)) \to 1, \ \nu(t) \to 0, \ \beta(t) \to \beta_* \quad \text{as } t \to \infty,$$

where $\beta_* := \left(\beta^2(0) + \nu^2(0) + \frac{2\nu^3(0)}{3}\right)^{1/2}$.

Theorem 3.2 immediately implies the following corollary about the asymptotics of the variance V_{est} in (2.4) for the predictive Student's t-distribution.

Corollary 3.1. *For any $\alpha(0), \beta(0), \nu(0) > 0$, we have*

$$\left(1 - \frac{\varkappa}{\alpha}\right) \frac{\alpha - A(\alpha)}{\alpha - 1} < \frac{V_{\text{est}}}{V} < \frac{\alpha - A(\alpha)}{\alpha - 1} \quad \text{for all large enough } t.$$

In particular,

$$V_{\text{est}} \to V \quad \text{as } t \to \infty.$$

Proof. From Theorem 3.2, item 1, we have by definition of S_*

$$\left(1 - \frac{\varkappa}{\alpha}\right)(\alpha - A(\alpha))V < \frac{\beta(\nu+1)}{\nu} < (\alpha - A(\alpha))V \quad \text{for all sufficiently large } t.$$

Deviding these inequalities by $\alpha - 1$ and recalling that $\alpha(t) \to \infty$ as $t \to \infty$ and $A(\alpha) \to 1$ as $\alpha \to \infty$, we obtain the desired result. □

Remark 3.3. One can show that $\varkappa/\alpha$ in the definition of the function $\sigma_\varkappa(\alpha)$ can be replaced by $\tilde{\varkappa}/\alpha^2$ with a sufficiently large $\tilde{\varkappa}$. In particular, the asymptotics in Corollary 3.1 will assume the form

$$\left|\frac{V_{\text{est}}}{V} - 1\right| = 1 + O(\alpha^{-2}) \quad \text{as} \quad t \to \infty.$$

The proof would require obtaining an extra term in the asymptotics of $A(\alpha)$ as $\alpha \to \infty$. However, we will not elaborate on these details.

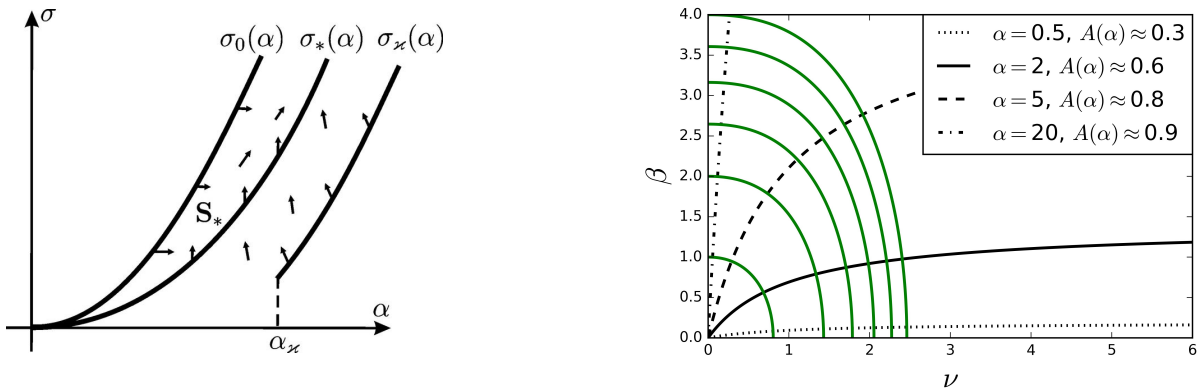

Figure 3.3: Left: The curves $\sigma_0(\alpha), \sigma_\varkappa(\alpha)$ given by (3.19), the curve $\sigma_*(\alpha)$ from Theorem 3.2, item 1, and the region $\mathbf{S}_*$ given by (3.20). The arrows indicate the directions of the vector field. Right: Green lines are the curves given by (3.21) for $C = 1, 4, 7, 10, 13, 16$. Black lines are the curves $\mathbf{C}_{\alpha,V}$ given by $\frac{\beta(\nu+1)}{\nu} = (\alpha - A(\alpha))V$ with $V = 1$. The black and green curves are orthogonal to each other.

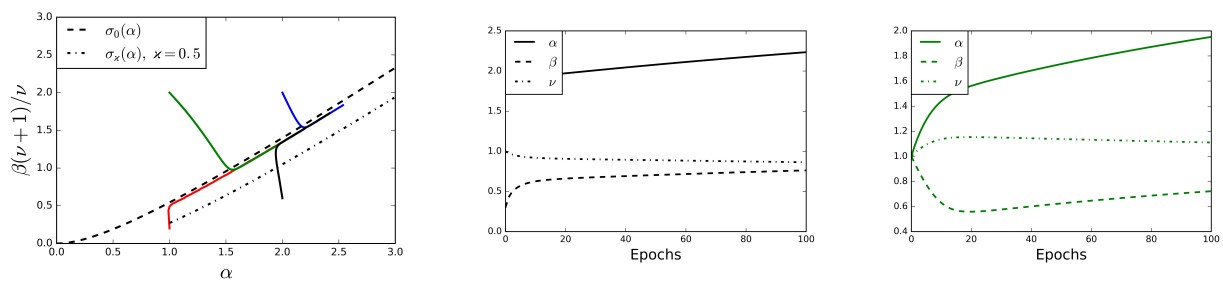

Figure 3.4: Left: Several trajectories obtained via iterating (3.2) for 2000 samples drawn from the normal distribution with mean 0 and variance 1. Middle/Right: Graphs of α, β, ν plotted versus the number of epochs, corresponding to the lower-right/upper-left trajectory in the left figure.

Remark 3.4. Suppose the number of observations tends to infinity. Then in the standard conjugate prior update (2.6), the parameters α, β, ν tend to infinity and the estimated mean and variance given by (2.4) converge to the ground truth mean $\mathbb{E}[\mathbf{y}]$ and variance $V = \mathbb{V}[\mathbf{y}]$, while

$$\mathbb{E}[\mu] \to \mathbb{E}[\mathbf{y}], \quad \mathbb{V}[\mu] \to 0, \quad \mathbb{E}[\tau] \to \frac{1}{\mathbb{V}[\mathbf{y}]}, \quad \mathbb{V}[\tau] \to 0.$$

The situation is quite different in Theorem 3.2. Although the parameter α tends to infinity, β converges to a finite positive value and ν converges to zero. Nevertheless, the estimated variance V_{est} in Corollary 3.1 converges to the ground truth variance $V = \mathbb{V}[\mathbf{y}]$, while (due to (2.2), (2.3), and (2.4))

$$\mathbb{E}[\mu] \to \mathbb{E}[\mathbf{y}], \quad \mathbb{V}[\mu] \to \mathbb{V}[\mathbf{y}], \quad \mathbb{E}[\tau] \to \infty, \quad \mathbb{V}[\tau] \to \infty.$$

3.4 Dynamics of α, β, ν. Proof of Theorem 3.2

First, we show that $\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right]$ and $\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]$ simultaneously vanish on the two-dimensional manifold

$$\left\{(\alpha, \beta, \nu) \in \mathbb{R}^3 : \alpha, \beta, \nu > 0 \text{ and } \frac{\beta(\nu+1)}{\nu} = \sigma_0(\alpha)\right\} \tag{3.23}$$

(the upper boundary of $\mathbf{S}_*$ in Fig. 3.3, left), where $\sigma_0(\alpha)$ is defined in (3.19). We will also see that $\dot{\sigma}, \dot{\beta} > 0$ and $\dot{\nu} < 0$ in $\mathbf{S}_*$.

Lemma 3.2. *We have*

$$\dot{\beta} = \dot{\nu} = 0 \quad \textit{if } \sigma = \sigma_0(\alpha), \tag{3.24}$$

$$\begin{aligned} &\dot{\sigma}, \dot{\beta} < 0, \ \dot{\nu} > 0 \quad \textit{if } \sigma > \sigma_0(\alpha), \\ &\dot{\sigma}, \dot{\beta} > 0, \ \dot{\nu} < 0 \quad \textit{if } \sigma < \sigma_0(\alpha). \end{aligned} \tag{3.25}$$

Proof. **1.** Let us show that $\dot{\nu} = 0$ if $\sigma = \sigma_0(\alpha)$. Due to (3.14),

$$\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]\Bigg|_{\sigma=\sigma_0(\alpha)} = \frac{1}{2\nu(\nu+1)}\left(\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\frac{(2\alpha+1)z^2}{2(\alpha-A)+z^2}e^{-\frac{z^2}{2}}dz - 1\right).$$

Now the equation $\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]\Big|_{\sigma=\sigma_0(\alpha)} = 0$ can be rewritten as follows:

$$\int_{-\infty}^{\infty}\frac{(2\alpha+1)z^2}{2(\alpha-A)+z^2}e^{-\frac{z^2}{2}}dz - 1 = 0, \tag{3.26}$$

which is equivalent to (3.15). By Lemma 3.1, it has a unique root $A = A(\alpha) \in (0, \min(\alpha, 1))$.

Since $\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]$ is decreasing with respect to σ due to (3.14), the assertions about ν in (3.25) follow.

2. Next, we show that $\dot{\beta} = 0$ if $\sigma = \sigma_0(\alpha)$. Due to (3.13), (3.15), and (3.16),

$$\begin{aligned} \mathbb{E}\left[\frac{\partial K}{\partial \beta}\right]\Bigg|_{\sigma=\sigma_0(\alpha)} &= \frac{\nu+1}{\nu V}\left(\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\frac{2\alpha+1}{2(\alpha-A)+z^2}e^{-\frac{z^2}{2}}dz - \frac{\alpha}{\alpha-A}\right) \\ &= \frac{\nu+1}{\nu V}\left((2\alpha+1)F(\alpha-A) - \frac{\alpha}{\alpha-A}\right) = 0. \end{aligned}$$

Since the expression in the brackets in (3.13) is increasing with respect to σ, the assertions about β in (3.25) follow. □

Now we show that the trajectories $(\nu(t), \beta(t))$ lie on curves that do not depend on α or V, see the green lines in Fig. 3.3 (right).

Lemma 3.3. *Let $\beta(t), \nu(t)$ $(t > 0)$ satisfy the last two equations in* (3.11) *(for an arbitrary $\alpha(t) > 0$). Then there is $C > 0$ such that all the points $(\nu(t), \beta(t))$ belong to the integral curve* (3.21) *of the equation* (3.22).

Proof. Using the equality $\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty} e^{-z^2/2}dz = 1$, we rewrite (3.14) as follows:

$$\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right] = -\frac{1}{\nu(\nu+1)}\left(\frac{\alpha+1/2}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\frac{1}{1+\frac{V}{\sigma}\frac{z^2}{2}}e^{-\frac{z^2}{2}}dz - \alpha\right).$$

Combining this relation with (3.13) shows that the points $(\nu(t), \beta(t))$ belong to the integral curves of the differential equation (3.22), because

$$\frac{d\beta}{d\nu} = \frac{\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right]}{\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]} = -\frac{\nu(\nu+1)}{\beta}$$

Separating variables in this equation, one can see that the integral curves are given by (3.21). □

Now we show that $\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right]$ is strictly negative on the manifold (3.23), and, hence, neither system (3.8), nor system (3.11) possesses an equilibrium.

Lemma 3.4. *We have*

$$\dot{\alpha} > 0 \quad \text{if } \sigma = \sigma_0(\alpha),\ \alpha > 0. \tag{3.27}$$

Moreover, for any $\varkappa > 0$, there exists $\alpha_\varkappa > 0$ such that

$$\dot{\alpha} < 0 \quad \text{if } \sigma = \sigma_\varkappa(\alpha),\ \alpha > \alpha_\varkappa. \tag{3.28}$$

Proof. **1.** Due to (3.12),

$$\begin{aligned}\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right]\Bigg|_{\sigma=\sigma_\varkappa(\alpha)} &= \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\ln\left(1+\frac{1}{\left(1-\frac{\varkappa}{\alpha^2}\right)(\alpha-A)}\frac{z^2}{2}\right)e^{-\frac{z^2}{2}}dz \\ &\quad + \Psi(\alpha) - \Psi\left(\alpha+\frac{1}{2}\right).\end{aligned} \tag{3.29}$$

Using that $\Psi(\alpha) - \Psi(\alpha+1/2) \to 0$ and $A(\alpha) \to 1$ as $\alpha \to \infty$, we see that

$$\lim_{\alpha\to\infty}\mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right]\Bigg|_{\sigma=\sigma_\varkappa(\alpha)} = 0. \tag{3.30}$$

2. To complete the proof of (3.27), it suffices (due to (3.30)) to show that the derivative of the right-hand side in (3.29) is positive for $\varkappa = 0$ and all $\alpha > 0$. We denote the derivative of the right-hand side in (3.29) by $G_\varkappa(\alpha)$. To calculate it, we set

$$B(\alpha) := 1 - \frac{\varkappa}{\alpha}, \quad \gamma(\alpha) = \frac{1}{2B(\alpha-A)}.$$

Then

$$\frac{\partial}{\partial\alpha}\ln(1+\gamma z^2) = -\frac{z^2}{1+\gamma z^2}\left(\frac{(1-A')}{2B(\alpha-A)^2} + \frac{B'}{2B^2(\alpha-A)}\right). \tag{3.31}$$

If $\varkappa = 0$, then $B = 1$, $B' = 0$, and (due to (A.1))

$$\frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{z^2}{1+\gamma z^2} = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{z^2}{1+\frac{1}{2(\alpha-A)} z^2} = \frac{2(\alpha - A)}{2\alpha+1}. \tag{3.32}$$

Using (3.31) and (3.32), we obtain

$$G_0(\alpha) = -\frac{1-A'}{(2\alpha+1)(\alpha-A)} + \Psi'(\alpha) - \Psi'\left(\alpha + \frac{1}{2}\right),$$

or, using Lemma 3.1 (item 4), equivalently,

$$G_0(\alpha) = -\frac{2}{(2\alpha+1)^2 A} + \Psi'(\alpha) - \Psi'\left(\alpha + \frac{1}{2}\right).$$

Due to the inequality $A(\alpha) > \frac{2\alpha}{2\alpha+3}$ (see the first inequality in Lemma 3.1, item 3),

$$G_0(\alpha) > \tilde{G}_0(\alpha),$$

where

$$\tilde{G}_0(\alpha) = -\frac{2\alpha+3}{(2\alpha+1)^2\alpha} + \Psi'(\alpha) - \Psi'\left(\alpha + \frac{1}{2}\right). \tag{3.33}$$

Therefore, for the proof of (3.27) it remains to show that

$$\tilde{G}_0(\alpha) > 0 \quad \text{for all } \alpha > 0. \tag{3.34}$$

2.1. First, we prove (3.34) for large α. Using the asymptotics [1, Sec. 6.4.12]

$$\Psi'(\alpha) - \Psi'\left(\alpha + \frac{1}{2}\right) = \frac{1}{2\alpha^2} + \frac{1}{4\alpha^3} + O\left(\frac{1}{\alpha^5}\right) \quad \text{as } \alpha \to \infty \tag{3.35}$$

we obtain from (3.33)

$$\tilde{G}_0(\alpha) = \frac{3}{8\alpha^4} + O\left(\frac{1}{\alpha^5}\right) > 0 \quad \text{for all sufficiently large } \alpha. \tag{3.36}$$

2.2. Due to (3.36), to complete the proof of (3.34) it now suffices to show that

$$\tilde{G}_0(\alpha) - \tilde{G}_0(\alpha+1) > 0 \quad \text{for all } \alpha > 0.$$

Applying the recurrence relation $\Psi'(z+1) = \Psi'(z) - 1/z^2$ (see [1, Sec. 6.4.6]), we obtain from (3.33)

$$\begin{aligned}\tilde{G}_0(\alpha) - \tilde{G}_0(\alpha+1) &= -\frac{2\alpha+3}{(2\alpha+1)^2\alpha} + \frac{2(\alpha+1)+3}{(2(\alpha+1)+1)^2(\alpha+1)} + \frac{1}{\alpha^2} - \frac{1}{(\alpha+1/2)^2} \\ &= \frac{3(4\alpha+3)}{(2\alpha+1)(\alpha+1)(2\alpha+3)^2\alpha^2} > 0,\end{aligned}$$

which proves (3.34) and thus completes the proof of (3.27).

3. Now consider the case $\varkappa > 0$. Note that the expression in the brackets in (3.31) is positive due to Lemma 3.1 and the fact that $B'(\alpha) > 0$. Hence, we obtain from (3.31)

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln(1+\gamma z^2) &= -2B(\alpha-A)\frac{z^2}{2B(\alpha-A)+z^2}\left(\frac{1-A'}{2B(\alpha-A)^2}+\frac{B'}{2B^2(\alpha-A)}\right)\\ &< -\frac{z^2}{2(\alpha-A)+z^2}\left(\frac{1-A'}{\alpha-A}+\frac{B'}{B}\right).\end{aligned}$$

Combining the latter inequality with (3.32) and using Lemma 3.1 (item 4), we have

$$G_\varkappa(\alpha) < -\frac{2}{(2\alpha+1)^2A} - \frac{B'}{B(2\alpha+1)} + \Psi'(\alpha) - \Psi'\left(\alpha+\frac{1}{2}\right).$$

Additionally using the asymptotics in (3.35) and the expansion of $A(\alpha)$ in (3.18) as $\alpha\to\infty$, we obtain

$$G_\varkappa(\alpha) < -\frac{\varkappa}{2\alpha^3} + o\left(\frac{1}{\alpha^3}\right) < 0 \quad \text{for all sufficiently large } \alpha.$$

□

Proof of Theorem 3.2. The arguments below are illustrated by Fig. 3.3 and can be separated into the following steps.

Proof of Item 1 Note that, for each fixed σ, the function $\mathbb{E}\left[\frac{\partial K}{\partial\alpha}\right]$ is monotonically decreasing with α. Furthermore by Lemma 3.4, we have $\dot\alpha > 0$ on $(\alpha, \sigma_0(\alpha))$. On the other hand, for large α, the asymptotic expansion of $\psi(\alpha)$ implies $\dot\alpha < 0$. Thus for each fixed σ there exists a unique value $\alpha_*(\sigma)$ such, that $\dot\alpha = 0$. Moreover, since $\mathbb{E}\left[\frac{\partial K}{\partial\alpha}\right]$ depends monotonically on σ and α, the function $\alpha_*(\sigma)$ is smooth and can be inverted to a function $\sigma_*(\alpha)$ by implicit function theorem. By construction, $\sigma_*(\alpha)$ satisfies all the properties in Theorem 3.2, items 1.

Proof of Item 2 We argue by contradiction. Suppose there does not exist t_0 such that $(\alpha(t_0), \beta(t_0), \nu(t_0)) \in \mathbf{S}_*$. Then by construction α has to decrease or increase monotonically, since it can never cross the curve $(\alpha, \sigma_*(\alpha)$. Suppose, that it decreases. Then, since $\sigma_*(\alpha)$ monotonically increases with α we know, that $\sigma(t)$ stays bounded. Furthermore by Lemma 3.2, also $\sigma(t)$ is monotonically increase. Thus, using compactness and monotonicity in time, there must exist $(\hat\alpha, \hat\sigma)$ such that $\lim_{t\to\infty}(\alpha(t), \sigma(t)) = (\hat\alpha, \hat\sigma)$. At that point the derivative of $\alpha(t)$ has to vanish, which implies, that $(\hat\alpha, \hat\sigma)$ has to lie on the curve $(\alpha, \sigma_*(\alpha))$. But on that curve the time-derivative of σ is bounded away from zero, which is a contradiction.

Proof of Item 3 Item 3 in Theorem 3.2 follows from Lemma 3.3.

Proof of Item 4 Let us prove item 4 in Theorem 3.2. Due to item 2, we can assume that $t_0 = 0$, so that $(\alpha(0), \beta(0), \nu(0)) \in \mathbf{S}_*$. By Lemma 3.2, $\alpha(t)$ and $\beta(t)$ are increasing, while $\nu(t)$ is decreasing. Furthermore, by Lemma 3.3, $\beta(t)$ is bounded for all t. Let us show that

$\nu(t) \to 0$ as $t \to \infty$. Since the vector field has singularities only for $\nu = 0$ or $\beta = 0$, it remains to exclude the following two cases.

Case 1: $\nu(t) \to \tilde{\nu}$ as $t \to \infty$ for some $\tilde{\nu} > 0$. In this case, $\beta(t) \to \tilde{\beta}$ for some finite $\tilde{\beta} > 0$, and hence $\alpha(t) \to \tilde{\alpha}$ for some finite $\tilde{\alpha} > 0$ since the trajectory must stay in $\mathbf{S}_*$. Therefore, $(\tilde{\alpha}, \tilde{\beta}, \tilde{\nu})$ must be an equilibrium of system (3.11). This contradicts Lemmas 3.2 and 3.4.

Case 2: $\nu(t) \to 0$ as $t \to T$ for some finite $T > 0$. In this case, $\beta(t) \to \tilde{\beta}$ as $t \to T$ for $\tilde{\beta} > 0$ and hence $\alpha(t) \to \infty$ as $t \to T$ since the trajectory must stay in $\mathbf{S}_*$. But this is possible only if $\dot{\alpha} = \mathbb{E}\left[\frac{\partial K}{\partial \alpha}\right] \to \infty$ as $\sigma, \alpha \to \infty$, which is not the case due to (3.12).

By having excluded Cases 1 and 2, we see that $\nu(t) \to 0$ as $t \to \infty$. Then $\beta(t) \to \tilde{\beta}$, where $\tilde{\beta} := \left(\beta^2(0) + \nu^2(0) + \frac{2\nu^3(0)}{3}\right)^{1/2}$, due to Lemma 3.3. Hence, $\frac{\beta(t)(\nu(t)+1)}{\nu(t)} \to \infty$ and $\alpha(t) \to \infty$ since the trajectory must stay in $\mathbf{S}_*$. Finally, by Lemma 3.1, $A(\alpha(t)) \to 1$. $\square$

4 Dynamics of m, β, ν for a fixed α.

4.1 Estimation of the variance. The curves of equilibria

According to Theorem 3.2, neither system (3.8), nor system (3.11) possesses an equilibrium. Moreover, the parameter α tends to infinity during learning and ν tends to zero. In this section, motivated by Remark 2.2, we fix α and update only m, β, ν in (3.2):

$$m_{\text{new}} := m - \lambda \frac{\partial K}{\partial m}, \quad \beta_{\text{new}} := \beta - \lambda \frac{\partial K}{\partial m}, \quad \nu_{\text{new}} := \nu - \lambda \frac{\partial K}{\partial \nu}. \tag{4.1}$$

As in Sec. 3, taking into account Theorem 3.1, we assume that the mean m has already been learned: $m = \mathbb{E}[\mathbf{y}]$ (Condition 3.2). Then the corresponding approximating dynamical system is given by the two equations for β, ν from (3.8):

$$\dot{\beta} = -\mathbb{E}\left[\frac{\partial K}{\partial \beta}\right], \quad \dot{\nu} = -\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right], \tag{4.2}$$

where the right-hand sides are explicitly given by (3.13) and (3.14). We consider this system on the quadrant $\{(\nu, \beta) \in \mathbb{R}^2 : \nu, \beta > 0\}$. Due to (3.25), this quadrant is forward invariant.

Remark 4.1. As in Remark 3.1, we conclude from (2.14) that system (4.2) defines the gradient flow with the potential $D_{\text{KL}}(q \| p_{\text{pred}}(\cdot; w))$, where $p_{\text{pred}}(\cdot; w))$ is the probability distribution of the Student's t-distribution $t_{2\alpha}(y | \mathbb{E}[\mathbf{y}], \beta(\nu+1)/(\nu\alpha))$.

Theorem 4.1. *Let $\alpha > 0$ be fixed. Then the following hold.*

1. *Dynamical system* (4.2) *possesses a globally attracting family of equilibria lying on the curve*

$$\mathbf{C}_{\alpha,V} := \left\{(\nu, \beta) \in \mathbb{R}^2 : \nu > 0, \ \frac{\beta(\nu+1)}{\nu} = \sigma_0(\alpha)\right\}, \tag{4.3}$$

where $\sigma_0(\alpha)$ is defined in (3.19).

2. *Each trajectory* $(\nu(t), \beta(t))$ *lies on one of the integral curves* (3.21). *If* $(\nu(0), \beta(0))$ *lies below the curve* $\mathbf{C}_{\alpha,V}$*, then* $\nu(t)$ *decreases and converges to* ν_* *and* $\beta(t)$ *increases and converges to* β_**. If* $(\nu(0), \beta(0))$ *lies above the curve* $\mathbf{C}_{\alpha,V}$*, then* $\nu(t)$ *increases and converges to* ν_* *and* $\beta(t)$ *decreases and converges to* β_**. In both cases,* (ν_*, β_*) *is the point of intersection of the corresponding integral curve and the curve of equilibria* $\mathbf{C}_{\alpha,V}$*. See Figure 3.3.*

3. *The family of integral curves* (3.21) *is orthogonal to the family of the curves of equilibria* $\{\mathbf{C}_{\alpha,V}\}_{\alpha,V>0}$*.*

Proof. Lemmas 3.2 and 3.3 imply items 1 and 2. Let us prove item 3. Assume that (ν, β) is a point of intersection of the curves

$$\begin{aligned} \beta = f(\nu) &:= (\alpha - A)V\frac{\nu}{\nu+1} && \text{(a curve of equilibria)}, \\ \beta = g(\nu) &:= \left(C - \nu^2 - \frac{2\nu^3}{3}\right)^{1/2} && \text{(an integral curve)}. \end{aligned} \tag{4.4}$$

Then $f'(\nu) = \frac{(\alpha-A)V}{(\nu+1)^2} = \frac{\beta}{\nu(\nu+1)}$. On the other hand, by Lemma 3.3, $g'(\nu) = -\frac{\nu(\nu+1)}{\beta}$. Thus, $f'(\nu)g'(\nu) = -1$, which implies the orthogonality of the two curves in (4.4). □

Theorem 4.1 immediately implies the following corollary.

Corollary 4.1. *Let* $\alpha > 0$ *be fixed. Then for any* $\beta(0), \nu(0) > 0$*, we have*

$$\tilde{V}_{\text{est}}(t) := \frac{\beta(t)(\nu(t)+1)}{(\alpha - A(\alpha))\nu(t)} \to V \quad \text{as } t \to \infty. \tag{4.5}$$

Figure 3.3 (right) shows the mutual configuration of the integral curves (3.21) and the curves of equilibria $\mathbf{C}_{\alpha,V}$ corresponding to different α. Figure 4.2 (left) shows several trajectories in the (ν, β) plane converging to the curve of equilibria $\mathbf{C}_{\alpha,V}$. Figure 4.2 (middle and right) shows that taking initial conditions with $m(0) \neq \mathbb{E}[\mathbf{y}]$ still yields the proper convergence of the estimated mean $m_{\text{est}} := m$ and the estimated variance $\tilde{V}_{\text{est}}$ in (4.5).

Remark 4.2. Due to Theorem 4.1, each trajectory $(\nu(t), \beta(t))$ of system (4.2) can be obtained by solving the scalar differential equation

$$\dot{\nu} = -\mathbb{E}\left[\frac{\partial K}{\partial \nu}\right]\Bigg|_{\beta=g(\nu)}, \tag{4.6}$$

where $g(\nu)$ is defined in (4.4) with a fixed $C > 0$ (uniquely determined by $\nu(0), \beta(0)$).

Furthermore, one can use other functions $\tilde{g}(\nu)$ in (4.6) instead of $g(\nu)$. Due to Lemma 3.24, the resulting ODE would still have an equilibrium ν_* such that $(\nu_*, \tilde{g}(\nu_*)) \in \mathbf{C}_{\alpha,V}$, and at this equilibrium, we would have

$$V = \frac{\beta_*(\nu_*+1)}{(\alpha - A(\alpha))\nu_*}, \quad \beta_* := \tilde{g}(\nu_*).$$

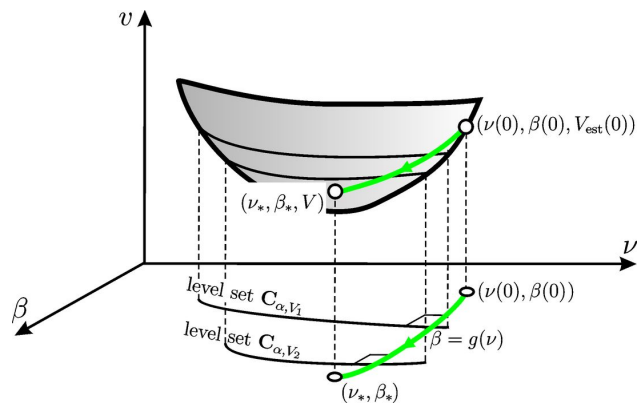

Figure 4.1: Schematic surface (4.7), its level sets $\mathbf{C}_{\alpha,V_1}$ and $\mathbf{C}_{\alpha,V_2}$, and the trajectory (in green) connecting the initial point $(\nu(0), \beta(0), V_{\text{est}}(0))$ and the target point (ν_*, β_*, V). The projection $\beta = g(\nu)$ of the trajectory is orthogonal to the level sets, i.e., the trajectory follows the gradient descent on the surface (4.7).

One can also show that this equilibrium is globally stable for a broad class of functions $\tilde{g}(\nu)$.

However, the function $g(\nu)$ from (4.4), corresponding to the integral curve (3.21), is *optimal* in the following sense, see Fig. 4.1. Consider the two-dimensional surface in $\mathbb{R}^3$

$$\left\{(\nu, \beta, v) \in \mathbb{R}^3 : v = \frac{\beta(\nu+1)}{(\alpha - A(\alpha))\nu}\right\} \tag{4.7}$$

(with α fixed). Then the initial point $(\nu(0), \beta(0), \tilde{V}_{\text{est}}(0))$ (where $\tilde{V}_{\text{est}}(t)$ is defined in (4.5)) and the target point (ν_*, β_*, V) (where ν_*, β_* are defined in Theorem 4.1, item 2) both lie on this surface. On the other hand, the curves $\{\mathbf{C}_{\alpha,V_1}\}_{V_1>0}$ are the level sets of this surface. Hence, due to Theorem 4.1, item 3, the curve $\beta = g(\nu)$ corresponds to the path of the *gradient descent* (or *ascent*) connecting the initial point $(\nu(0), \beta(0), V_{\text{est}}(0))$ and the target point (ν_*, β_*, V).

Remark 4.3. It follows from Remark 4.1 and item 3 in Theorem 4.1 that, for any fixed α, V, the curve $\mathbf{C}_{\alpha,V}$ is the set of minima of the potential $D_{\text{KL}}(q \| p_{\text{pred}}(\cdot; w))$, while all the other curves $\mathbf{C}_{\alpha_1,V_1}$, $\alpha_1, V_1 > 0$, are the level sets of this potential.

Remark 4.4. The situation in Theorem 4.1 is different both from the standard CP update (2.6) and from the GCP update (3.2) (cf. Remark 3.4). First, the parameter α is now fixed. Furthermore, each trajectory of system (4.2) (approximating the GCP update (4.1)) converges to a finite equilibrium (ν_*, β_*), where $\nu_*, \beta_* > 0$. Nevertheless, the estimated variance $\tilde{V}_{\text{est}}$ given by (4.5) again converges to the ground truth variance $V = \mathbb{V}[\mathbf{y}]$.

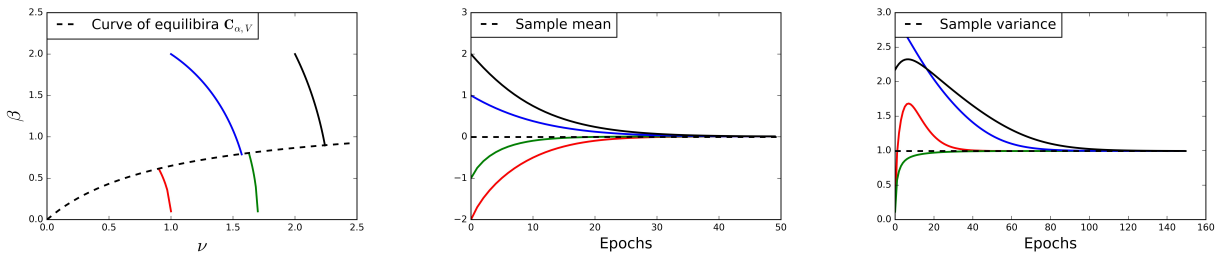

Figure 4.2: Several trajectories obtained via iterating (4.1) with 2000 samples drawn from the normal distribution with mean 0 and variance 1. The parameter α is fixed: $\alpha = 2$, $A(\alpha) \approx 0.619$. Left: plane (ν, β), with the initial condition $m(0) = 0$ for all trajectories. Middle and Right: means m and variances $\tilde{V}_{\text{est}}$ plotted versus the number of epochs. The initial conditions for β and ν are the same as for the trajectories of the respective colors in the left plot, while $m(0) = -2, -1, 1, 2$.

5 Role of a fixed α

5.1 Sensitivity to outliers

It is well known that outliers essentially influence the estimate of the mean m if one uses the standard squared error loss

$$\mathcal{L}_{\text{SE}}(\overline{y}, m) = (m - \overline{y})^2.$$

The same is true when one estimates both mean m and precision p (inverse variance) via maximizing the log-likelihood of a normal distribution, or, equivalently, minimizing the loss

$$\mathcal{L}_{\text{ML}}(\overline{y}, m, p) = p(m - \overline{y})^2 - \ln p.$$

The reason is that, in both cases, the derivatives of the loss functions $\mathcal{L}_{\text{SE}}$ and $\mathcal{L}_{\text{ML}}$ with respect to m are proportional to $m - \overline{y}$, while the derivative of $\mathcal{L}_{\text{ML}}(\overline{y}, m, p)$ with respect to p contains even $(m - \overline{y})^2$. It turns out that the GCP update (4.1) is much less sensitive to outliers, see Fig. 5.1. This can be explained by the fact that the derivatives of the KL divergence with respect to m, β and ν are bounded with respect to $m - \overline{y}$, see (3.4), (3.6), and (3.7). Moreover, $\frac{\partial K}{\partial m}$ even vanishes as $m - \overline{y} \to \infty$. Another explanation is that the GCP update is equivalent to maximizing the likelihood of the Student's t-distribution (item 1 in Remark 2.1). It is known [21] that the optimal value of m is different from the sample mean due to downweighting the outlying observations.

5.2 Learning speed in clean and noisy regions

5.2.1 Observations

When one approximates the parameters m, β, ν by deep neural networks, one represents these parameters as functions of an input variable $x \in X$ (where $X \subset \mathbb{R}^N$) and of a set of weights $w \in \mathbb{R}^M$. Since neural networks have finite capacity (M is finite), they cannot perfectly

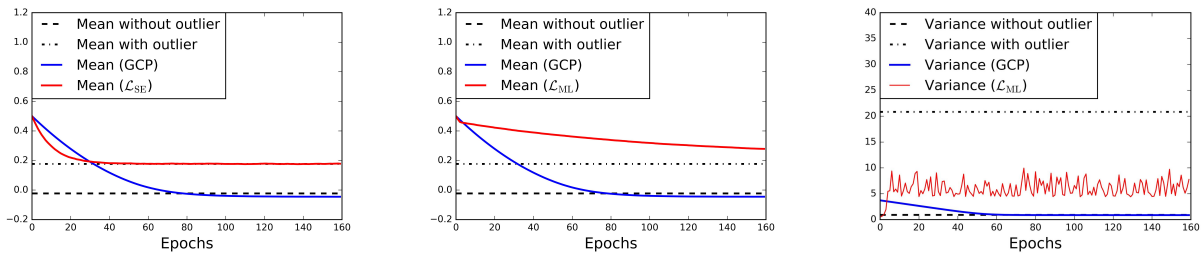

Figure 5.1: Fitting 500 samples drawn from the normal distribution with mean 0 and variance 1 and supplemented by an outlier $\overline{y} = 100$. Left: Means fitted with the GCP update (4.1) and, respectively, with the standard squared error loss $\mathcal{L}_{\mathrm{SE}}$. Middle/Right: Means/variances fitted with the GCP update (4.1) and, respectively, via maximizing the likelihood, i.e., minimizing $\mathcal{L}_{\mathrm{ML}}$. For the GCP update, the parameter α is fixed: $\alpha = 1$, $A(\alpha) \approx 0.46$, and the variance is estimated by $\tilde{V}_{\mathrm{est}}$ in (4.5).

approximate m, β, ν for all x simultaneously. Therefore, it is important to understand in which regions of the input space X the parameters are approximated better and in which worse, cf. [5]. This is directly related to the values of the gradients in the GCP update (4.1), which determine the learning speed. The faster the learning in a certain region occurs, the more influential this region is. In particular, we are interested in the learning speed in so called clean regions (where V is small) compared with noisy regions (where V is large).

Below, we will concentrate on the regime where the learning process starts and the initial conditions for β and ν satisfy

$$\beta(0) \approx \nu(0) \approx 1. \tag{5.1}$$

This is often the case if β and ν are approximated by neural networks with the softplus output, e.g.,

$$\beta = \ln\left(1 + e^{w}\right), \tag{5.2}$$

where $w \in \mathbb{R}$ is the input of the softplus output.

In the observations below, we denote the learning speed of the mean and the variance by $\mathrm{LerSp}(m)$ and $\mathrm{LerSp}(\mathrm{Var})$, respectively.

Observation 5.1. *Let α be small.*

1. *In clean regions (small V):* $\mathrm{LerSp}(m)$ *is of order* 1 *and* $\mathrm{LerSp}(\mathrm{Var})$ *is of order* 1.
2. *In noisy regions (large V):* $\mathrm{LerSp}(m)$ *is of order* $1/V$ *and* $\mathrm{LerSp}(\mathrm{Var})$ *is of order* α.

Observation 5.2. *Let α be large.*

1. *In clean regions (small V):* $\mathrm{LerSp}(m)$ *is of order* α *and* $\mathrm{LerSp}(\mathrm{Var})$ *is of order* 1.
2. *In noisy regions (large V):* $\mathrm{LerSp}(m)$ *is of order* $1/V$ *and* $\mathrm{LerSp}(\mathrm{Var})$ *is of order* α.

α	LerSp(m)	LerSp(Var)
Small	cl. > noisy	cl. > noisy
Large	cl. $\gg$ noisy	cl. < noisy

Table 5.1: Relative learning speed of estimated mean m and variance $\tilde{V}_{\text{est}}$ for clean (cl.) and noisy regions. Notation "<" and ">" stands for a "lower" and a "higher" speed, and "$\gg$" for a "much higher" speed.

Observations 5.1 and 5.2 are summarized in Table 5.1. In particular, we see that the mean is always learned faster in clean regions. Taking large α further increases the learning speed of the mean in clean regions, but simultaneously increases the learning speed of variance in noisy regions compared with clean regions.

Observation 5.3. *The values of β_* and ν_* to which the trajectory of* (4.2) *will converge are determined by the value $(\alpha - A(\alpha))V$.*

1. *If $(\alpha - A(\alpha))V \ll 1$, then $\beta_* \approx 0$ and $\nu_* \approx 1$.*
2. *If $(\alpha - A(\alpha))V \gg 1$, then $\beta_* \approx 1$ and $\nu_* \approx 0$.*

Small values of β and ν will lead to large gradients $\frac{\partial K}{\partial \beta}$ and $\frac{\partial K}{\partial \nu}$, respectively, which may cause large oscillations of $\tilde{V}_{\text{est}}$.

Observations 5.1–5.3 are illustrated in Fig. 5.2 and explained in detail below.

5.2.2 Justification of the observations

1. First, we analyze LerSp(Var). Consider the limit $V \to 0$ (clean regions). Due to (4.3), the point (ν, β) lies above the line of equilibria $\mathbf{C}_{\alpha,V}$ at a vertical distance of order 1 from it. Using (3.13) and (3.14), we see that

$$\lim_{V\to 0} \mathbb{E}\left[\frac{\partial K}{\partial \beta}\right] = \frac{1}{2\beta}, \quad \lim_{V\to 0} \mathbb{E}\left[\frac{\partial K}{\partial \nu}\right] = -\frac{1}{2\nu(\nu+1)},$$

i.e. (ν, β) approaches $\mathbf{C}_{\alpha,V}$ with speed of order 1.

Now consider the limit $V \to \infty$. Due to (4.3), the point (ν, β) lies below and to the right from the curve of equilibria $\mathbf{C}_{\alpha,V}$ at a horizontal distance of order 1 from it. Using (3.13) and (3.14), we see that

$$\lim_{V\to \infty} \mathbb{E}\left[\frac{\partial K}{\partial \beta}\right] = -\frac{\alpha}{\beta}, \quad \lim_{V\to \infty} \mathbb{E}\left[\frac{\partial K}{\partial \nu}\right] = \frac{\alpha}{\nu(\nu+1)},$$

i.e. (ν, β) approaches $\mathbf{C}_{\alpha,V}$ with speed of order α.

These arguments justify the assertions about LerSp(Var) in Observations 5.1 and 5.2.

Further, recall that the trajectory (ν, β) lies on one of the curves (3.21). Thus, if $(\alpha - A(\alpha))V \ll 1$ and (ν, β) approaches $\mathbf{C}_{\alpha,V}$, the value of β will approach 0, while ν will stay of

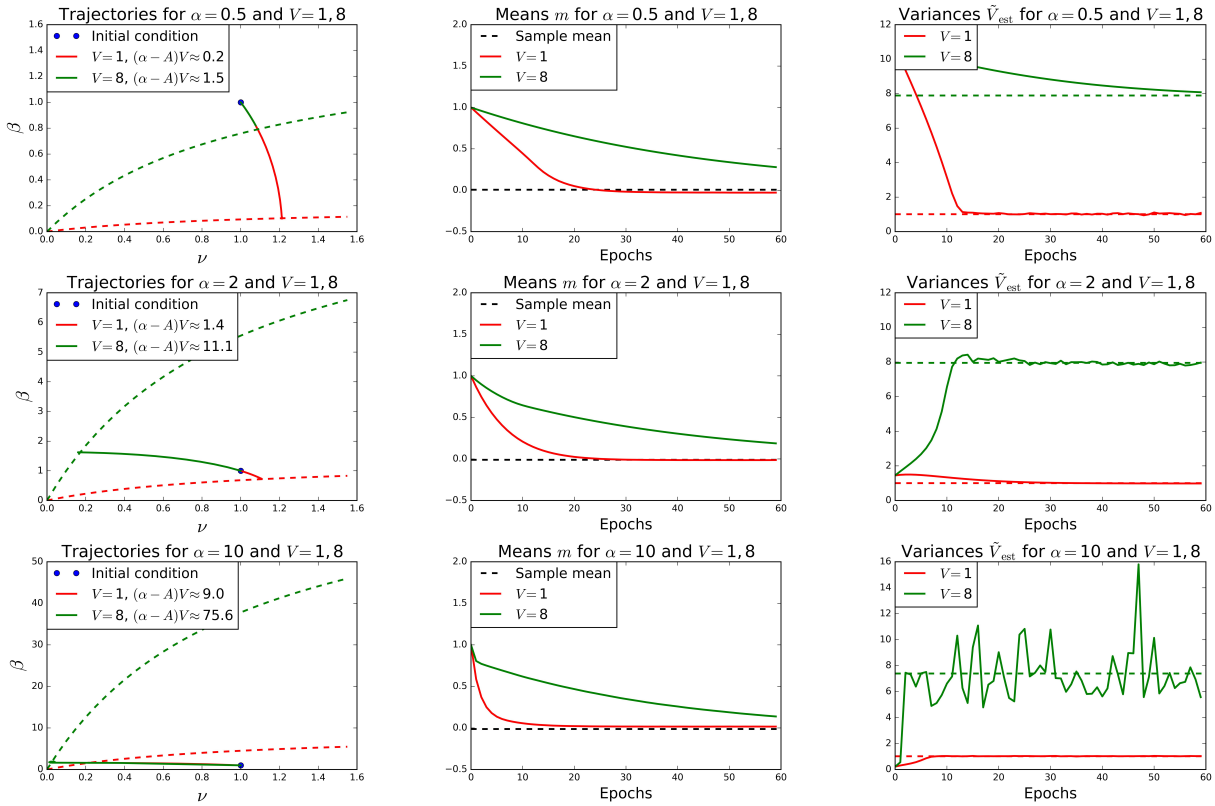

Figure 5.2: Trajectories (left) and graphs of means (middle) and variances (right) versus the number of epochs for different values of α and V, based on the GCP update (4.1) for 2000 samples drawn from the normal distribution with mean 0 and variance V. The initial conditions are $m(0) = \beta(0) = \nu(0) = 1$. The dashed lines in the left-hand column indicate the corresponding curves of equilibria $\mathbf{C}_{\alpha,V}$. The dashed lines in the right-hand column indicate the corresponding sample variances.

order 1. On the other hand, if $(\alpha - A(\alpha))V \gg 1$ and (ν, β) approaches $\mathbf{C}_{\alpha,V}$, the value of β will stay of order 1 and ν will approach 0. This is illustrated in Fig. 5.2 (left-hand column). This justifies Observation 5.3.

2. Now we analyze LerSp(m). Here we assume that the *variance V has already been estimated approximately.* We express this fact by assuming that the parameters β and ν are such that

$$\tilde{V}_{\text{est}} = cV \tag{5.3}$$

for some $c \in (c_1, c_2)$, where $c_2 > c_1 > 0$ do not depend on V. Without loss of generality,

assume that $\mathbb{E}[\mathbf{y}] = 0$ and $m > 0$. Then, due to (3.9),

$$\begin{aligned}\mathbb{E}\left[\frac{\partial K}{\partial m}\right] &= (\alpha + 1/2)\frac{1}{(2\pi)^{1/2}}\int\limits_{-\infty}^{\infty}\frac{m - V^{1/2}z}{c(\alpha - A)V + \frac{\left(m - V^{1/2}z\right)^2}{2}}e^{-\frac{z^2}{2}}dz \\ &= (\alpha + 1/2)\frac{1}{(2\pi V)^{1/2}}\int\limits_{-\infty}^{\infty}\frac{\frac{m}{V^{1/2}} - z}{c(\alpha - A) + \frac{\left(\frac{m}{V^{1/2}} - z\right)^2}{2}}e^{-\frac{z^2}{2}}dz.\end{aligned} \tag{5.4}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\frac{\partial K}{\partial m}\right] &= \frac{2\alpha + 1}{m} + o(V) \quad \text{as } V \to 0, \\ \mathbb{E}\left[\frac{\partial K}{\partial m}\right] &= \frac{k(\alpha)}{V} + o\left(\frac{1}{V}\right) \quad \text{as } V \to \infty,\end{aligned} \tag{5.5}$$

where

$$k(\alpha) = \frac{(\alpha + 1/2)m}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\frac{c(\alpha - A) - z^2/2}{(c(\alpha - A) + z^2/2)^2}e^{-\frac{z^2}{2}}dz.$$

The constant $k(\alpha)$ can be obtained by dividing the integral in the right-hand side of (5.4) by $V^{-1/2}$ and applying L'Hospital's rule. Using the properties of $A(\alpha)$ in (3.18), one can show that $k(\alpha)$ is positive, bounded, and bounded away from zero for all $\alpha > 0$.

These arguments justify the assertions about LerSp(m) in Observations 5.1 and 5.2.

6 GCP neural networks

6.1 Methods, architectures, and measures

Methods. We compare the GCP networks with the following:

1. the maximum likelihood method (ML), in which one maximizes the likelihood of the normal distribution,

2. the probabilistic back propagation (PBP) [7], which is one of the state-of-the-art Bayesian methods,

Architectures. We use 1-hidden layer networks for the parameters of the prior (2.15) with ReLU nonlinearities. Each network contains 50 hidden units for all the data sets below, except for the largest MSD set. For the latter, we use 100 hidden units. Our approach is directly applicable to neural networks of any depth and structure, however we kept one hidden layer for the compatibility of our validation with [5, 7, 14, 15].

Measures. We use two measures to estimate the quality of the fit.

1. The overall root mean squared error (*RMSE*).

Synthetic data		
	RMSE	AUC
ML	0.52	0.39
PBP	0.61	0.56
GCP	**0.32**	**0.25**

Table 6.1: RMSE and AUC for the synthetic data set. The values are given for a test set without outliers

2. The area under the following curve (AUC), measuring the trade-off between properly learning the mean and the variance. Assume the test set contains N samples. We order them with respect to their predicted variance. For each $n = 0, \ldots, N-1$, we remove n samples with the highest variance and calculate the RMSE for the remaining $N-n$ samples (with the lowest variance). We denote it by RMSE(n) and plot it versus n as a continuous piecewise linear curve. The second measure is the area under this curve normalized by $N-1$:

$$\mathrm{AUC} := \frac{1}{N-1} \sum_{n=0}^{N-2} \frac{\mathrm{RMSE}(n) + \mathrm{RMSE}(n+1)}{2}.$$

6.2 Synthetic data set

We generate a synthetic data set containing 2% of outliers. To do so, we chose the set X consisting of 400 points uniformly distributed on the interval $(-1, 1)$. For each $x \in X$, with probability 0.98 we sample y from the normal distribution with mean $\sin(3x)$ and standard deviation $0.5\cos^4 x$, and with probability 0.02 we sample y from a uniform distribution on the interval $(-4, 16)$. Figure 6.1 shows the data and the fits of ML, PBP, and GCP, together with the curves RMSE(n). We see that the mean predicted by GCP is much less affected by the outliers compared to that of ML and PBP. Further, we see that the standard deviation predicted by GCP according to (2.16) (in solid blue) is also less affected by the outliers. However, as we mentioned in Remark 2.2, the outliers prevent α from going to infinity, which is reflected in still higher values of the predicted standard deviation compared to the ground truth $0.5\cos^4 x$. Using the corrected variance $\tilde{V}_{\text{est}}$ given by (2.17) (in dashed blue) allows one to recover the ground truth standard deviation.

Table 6.1 shows the values of RMSE and AUC on the test set containing no outliers.

6.3 Real world data sets

We analyze the following publicly available data sets: Boston House Prices [6] (506 samples, 13 features), Concrete Compressive Strength [32] (1030 samples, 8 features), Combined Cycle Power Plant [11, 28] (9568 samples, 4 features), Yacht Hydrodynamics [4, 24] (308 samples, 6 features), Kinematics of an 8 Link Robot Arm Kin8Nm[1] (8192 samples, 8 feature), and

[1] http://mldata.org/repository/data/viewslug/regression-datasets-kin8nm/

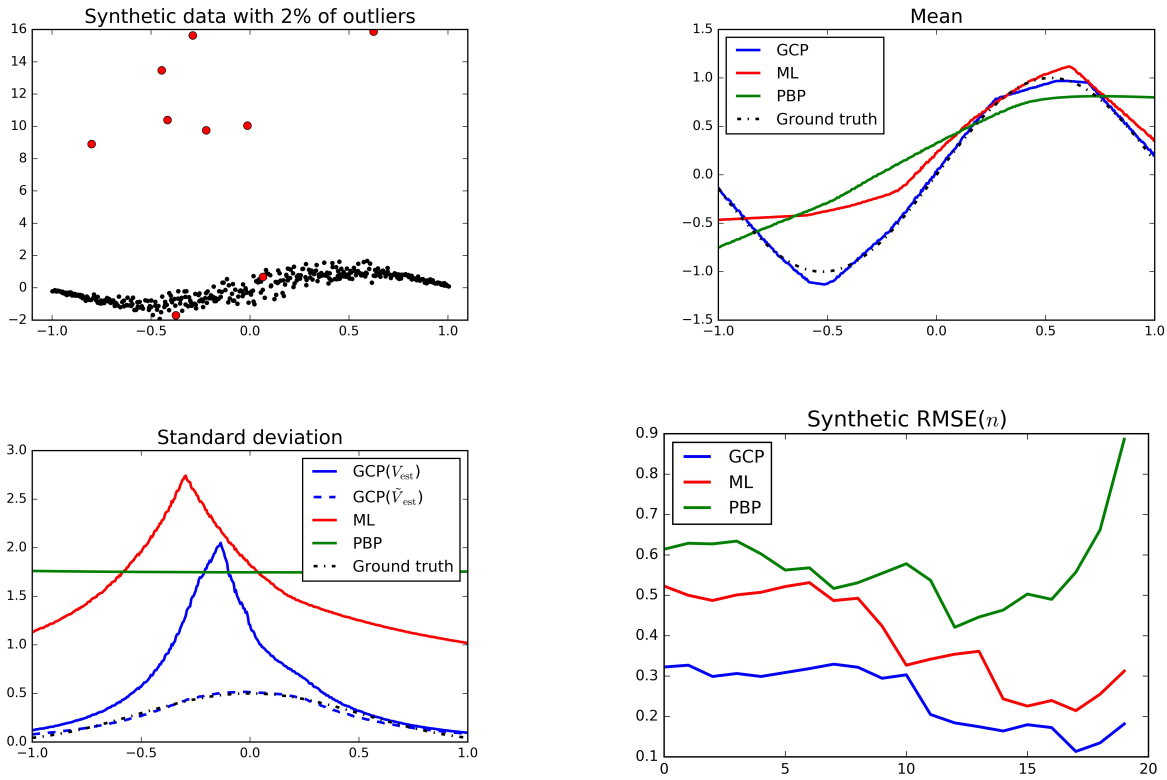

Figure 6.1: Top left: Synthetic data with mean $\sin(3x)$ and standard deviation $0.5\cos^4 x$ (black dots), complemented by 2% of outliers sampled from a uniform distribution on the interval $(-4, 16)$ (red disks). Top right: The ground truth mean $\sin(3x)$ and the means predicted by ML, PBP, and GCP. Bottom left: The ground truth standard deviation $0.5\cos^4 x$ and the standard deviations predicted by ML, PBP, and GCP via V_{est} in (2.16) and via the corrected formula for $\tilde{V}_{\text{est}}$ in (2.17). Bottom right: The curves RMSE(n) for the different methods.

Year Prediction MSD [16] (515345 samples, 90 features). Each data set, except for the year prediction MSD, is randomly split into 50 train-test folds with 95% of samples in each train subset. All the measure values reported below are the averages of the respective measure values over 50 folds. For the year prediction MSD, we used a single split recommended in [16]. The data sets are normalized so that the input features and the targets have zero mean and unit variance in the training set.

Tables 6.2 and 6.3 show the measure values of GCP in comparison with the other methods[2]. The values with the best mean and the values that are not significantly different from those with the best mean (due to the two-tailed paired difference test with $p = 0.05$) are marked in bold. We see that GCP achieves the best AUC values on all the data sets except

[2]We were not able to fit PBP for the largest MSD data set. In Table 6.3, we report the RMSE value for the latter based on [7].

	Boston		**Concrete**		**Power**	
	RMSE	AUC	RMSE	AUC	RMSE	AUC
ML	**3.79±1.44**	2.40±0.59	**5.60±0.63**	4.00±0.59	**4.09±0.31**	**3.75±0.24**
PBP	**3.54±1.29**	2.28±0.47	**5.58±0.60**	4.66±0.64	**4.09±0.26**	3.86±0.26
GCP	**3.63±1.57**	**1.88±0.44**	**5.56±0.69**	**3.59±0.47**	**4.14±0.31**	**3.64±0.36**

Table 6.2: RMSE and AUC for the Boston, Concrete, and MSD data sets.

	Yacht		**Kin8nm**		**MSD**	
	RMSE	AUC	RMSE	AUC	RMSE	AUC
ML	**0.76±0.38**	**0.25±0.08**	**0.09±0.01**	0.07±0.00	9.09±NA	5.40±NA
PBP	1.09±0.35	0.64±0.19	0.10±0.01	0.08±0.00	**8.88±NA**	NA
GCP	0.94±0.47	**0.27±0.12**	0.09±0.01	**0.06±0.00**	9.18±0.00	**5.30±NA**

Table 6.3: RMSE and AUC for the Yacht, Kin8nm, and MSD data sets.

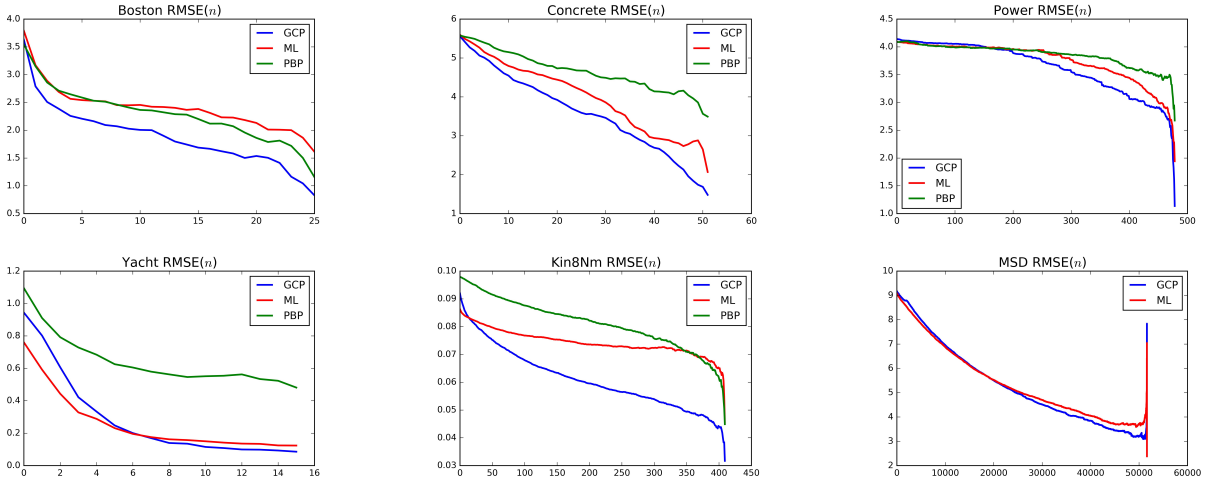

Figure 6.2: The curves RMSE(n) for the different methods and data sets from Tables 6.2 and 6.3.

for Yacht, on which its AUC is not significantly different from the best. Furthermore, it achieves the best RMSE value for the Boston data set and the values that are not significantly different from the best for Concrete and Power. Its RMSE values for Yacht, Kin8nm, and MSD are not the best, but still competitive.

Figure 6.2 shows the curves RMSE(n) for the different methods and data sets from Tables 6.2 and 6.3. We see that the curve RMSE(n) typically decays faster for the GCP compared to the other method.

7 Conclusion

Our goal was to learn a parametrically given ground truth probability distribution by artificial neural networks. For the unknown parameters, we introduced a prior distribution, whose parameters are to be learned by neural networks. In such a setting, one cannot directly update the prior's parameters by the Bayesian rule, but one should rather update the network's weights. Hence, we proposed to replace a full Bayesian update of prior's parameters by one gradient descent step in the direction of minimizing the KL divergence from the prior to the posterior distribution, which we called the GCP update. We showed that the GCP update is equivalent to the gradient ascent step that maximizes the likelihood of the predictive distribution. Interestingly, this result holds in general, independently of whether the posterior and prior distributions belong to the same family or not.

Next, we concentrated on the case where the ground truth distribution is normal with unknown mean and variance. A natural choice for the prior is the normal-gamma distribution. We obtained a dynamical system for its parameters that approximates the corresponding GCP update and analyzed it in detail. It revealed the convergence of the prior's parameters that is quite different from that for the standard Bayesian update, although in both cases the predictive Student's t-distribution converges to the ground truth normal distribution.

Furthermore, we analyzed how the GCP interacts with outliers in the training set (a small percentage of observations that do not come from the ground truth normal distribution). In the presence of outliers, the prior's parameter α (half the number of degrees of freedom of the predictive Student's t-distribution) does not tend to infinity any more. On one hand, this allows for a much better estimate of the mean of the ground truth normal distribution, compared with the ML method. On the other hand, this leads to overestimation of the variance of the ground truth distribution. We obtained, for the first time, an explicit formula that allows one to correct the estimate of the variance and recover the ground truth variance of the normal distribution.

Finally, we validated the GCP neural network on six real-world data sets and compared it with the ML and PBP neural networks. We measured the trade-off between properly learning the mean and the variance (reflected in the AUC values) and the overall error (RMSE). For all the data sets, the GCP demonstrated the best AUC values and either superior or competitive values of RMSE.

To conclude, we indicate several directions of future research:

1. A rigorous mathematical analysis of the influence of outliers on the dynamics of the prior's parameters seems to be feasible. One can relate the percentage of the outliers and a type of distribution they come from with the dynamical system (3.8), in which the expectations will be taken with respect to the new distribution (mixture of normal and the one from which the outliers are sampled).

2. Section 4 shows that one can fix α and still recover the ground truth normal distribution, while Sec. 5 indicates how different values of α may influence the learning speed in clean and noisy regions. The influence of α on the fit of the GCP neural networks for real-world data sets would be an interesting practical question. Our preliminary

analysis showed that fixing large α was beneficial for the largest MSD data set. For example, fixing $\alpha = 30$ yielded RMSE $= 8.89$ and AUC $= 5.13$ (cf. Table 6.3).

3. It is worth checking the GCP networks for other choices of ground truth and prior distributions.

4. The use of ensembles of MLs (called deep ensembles) was recently proposed in [14,15]. Since the fit of the GCP is typically better than that of the ML (cf. Tables 6.1, 6.2, and 6.3 and Figures 6.1 and 6.2), it is worth studying ensembles of GCPs and comparing them with ensembles of MLs.

A Properties of the function $A(\alpha)$: proof of Lemma 3.1

For $\alpha > 0$ and $A \in [0, \alpha]$, we study equation (3.15), which is equivalent to the following:

$$E(\alpha, A) := \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)z^2}{2(\alpha - A) + z^2} e^{-\frac{z^2}{2}} dz - 1 = 0. \tag{A.1}$$

Lemma A.1. *For each $\alpha > 0$, equation* (A.1) *has a unique root $A = A(\alpha) \in (0, \alpha)$. Furthermore,*

$$A(\alpha) = \alpha - \frac{4}{\pi}\alpha^2 + o(\alpha^2) \text{ as } \alpha \to 0.$$

Proof. **1.** Note that $E(\alpha, A)$ is increasing with respect to $A \in (0, \alpha)$ and $E(\alpha, \alpha) = 2\alpha > 0$. Hence, it remains to show that $E(\alpha, 0) < 0$. We have

$$E(\alpha, 0) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)z^2}{2\alpha + z^2} e^{-\frac{z^2}{2}} dz - 1 = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{z^2}{2\alpha + z^2}(1 - z^2) e^{-\frac{z^2}{2}} dz. \tag{A.2}$$

Now the inequality $E(\alpha, 0) < 0$ follows from (A.2) and the monotonicity of $\frac{z^2}{2\alpha+z^2}$. Indeed,

$$\begin{aligned}\int_0^1 \frac{z^2}{2\alpha + z^2}(1 - z^2) e^{-\frac{z^2}{2}} dz &< \int_0^1 \frac{1}{2\alpha + 1}(1 - z^2) e^{-\frac{z^2}{2}} dz \\ &= \int_1^{\infty} \frac{1}{2\alpha + 1}(z^2 - 1) e^{-\frac{z^2}{2}} dz < \int_1^{\infty} \frac{z^2}{2\alpha + z^2}(z^2 - 1) e^{-\frac{z^2}{2}} dz,\end{aligned}$$

where we have used the equality

$$\int_0^{\infty} e^{-\frac{z^2}{2}} dz = \int_0^{\infty} z^2 e^{-\frac{z^2}{2}} dz = \left(\frac{\pi}{2}\right)^{1/2}.$$

2. Now we prove the asymptotics of $A(\alpha)$. Using the function $F(x)$ defined in (3.16), we rewrite equation (A.1) in the form

$$(2\alpha + 1)(\alpha - A)F(\alpha - A) - \alpha = 0. \tag{A.3}$$

Using the expansion of $\operatorname{erfc}(x)$ around 0 (see [1, Sec. 7.1.6]) and formula (3.16), we have for all $x > 0$

$$xF(x) = \frac{\pi^{1/2} + \delta}{2} x^{1/2}, \quad \delta = o(1) \text{ as } x \to 0. \tag{A.4}$$

Now, for each $\alpha > 0$, we represent $A = \alpha - k^2\alpha^2$ and prove that $k = \frac{2}{\pi^{1/2}} + o(1)$ as $\alpha \to 0$. Combining the representation of A with (A.3) and (A.4), we obtain

$$\left(-1 + \frac{k\pi^{1/2}}{2} + \frac{k\delta}{2}\right) + k(\delta + \pi^{1/2})\alpha = 0.$$

Obviously, if $\alpha = \delta = 0$, we have $k = \frac{2}{\pi^{1/2}}$. Hence, by the implicit function theorem, $k = \frac{2}{\pi^{1/2}} + o(1)$ as $\alpha, \delta \to 0$. Recalling that $\delta = o(1)$ as $\alpha \to 0$, we complete the proof. □

Lemma A.2. *For each $\alpha \geq 1$, equation* (A.1) *has a unique root $A = A(\alpha) \in (0, 1)$. Furthermore,*

$$A(\alpha) = 1 - \frac{3}{2\alpha} + o\left(\frac{1}{\alpha}\right) \text{ as } \alpha \to \infty.$$

Proof. **1.** In the proof of Lemma A.1, we have shown that $E(\alpha, 0) < 0$. Due to the monotonicity of $E(\alpha, A)$ with respect to $A \in (0, \alpha)$, it remains to show that $E(\alpha, 1) > 0$. Using that $2\alpha - 2 \geq 0$, we have

$$\begin{aligned}
E(\alpha, 1) &= \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)z^2}{2\alpha - 2 + z^2} e^{-\frac{z^2}{2}} dz - 1 = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{z^2(3 - z^2)}{2\alpha - 2 + z^2} e^{-\frac{z^2}{2}} dz \\
&= \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \int_0^{\infty} z^2(3 - z^2) e^{-z^2/2 - (2\alpha - 2 + z^2)\xi} d\xi \, dz \\
&= \frac{1}{(2\pi)^{1/2}} \int_0^{\infty} e^{-(2\alpha-2)\xi} d\xi \int_{-\infty}^{\infty} z^2(3 - z^2) e^{-(\xi + 1/2)z^2} dz \\
&= 3 \int_0^{\infty} \exp(-(2\alpha - 2)\xi) \left(\frac{1}{(2\xi + 1)^{3/2}} - \frac{1}{(2\xi + 1)^{5/2}} \right) d\xi > 0.
\end{aligned}$$

2. Now we prove the asymptotics of $A(\alpha)$. Using the expansion of $\operatorname{erfc}(x)$ around ∞ (see [1, Sec. 7.1.23]) and formula (3.16), we have for all $x > 0$

$$xF(x) = \frac{1}{2} - \frac{1}{4x} + \frac{3}{8x^2} - \frac{15 + \delta}{16x^3}, \quad \delta = o(1) \text{ as } x \to \infty. \tag{A.5}$$

Now, for each $\alpha > 0$, we representing $A = 1 - k/\alpha$ and prove that $k = 3/2 + o(1)$ as $\alpha \to \infty$. Combining the representation of A with (A.3) and (A.5), we obtain

$$\frac{-12 + 38kz^2 - 28k^2z^3 + 8k^3z^4 + 8k - 28kz - 33z - 2\delta + 16k^2z^2 - z\delta}{(1 - z + kz^2)^3} = 0, \quad z := \frac{1}{\alpha}.$$

Obviously, if $\alpha = \delta = 0$, we have $k = 3/2$. Hence, by the implicit function theorem, $k = 3/2 + o(1)$ as $z, \delta \to 0$. Recalling that $\delta = o(1)$ as $z \to 0$, we complete the proof. □

Lemma A.3. *The function $A(\alpha)$ satisfies the differential equation in* (3.17).

Proof. Denoting by E_α and E_A the partial derivatives of the function $E(\alpha, A)$ with respect to α and A, respectively, and using the implicit function theorem, we have

$$A' = -E_A^{-1} E_\alpha. \tag{A.6}$$

Since $A(\alpha)$ satisfies the equation in (A.1), we obtain

$$\begin{aligned} E_\alpha &= -2\frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)z^2}{(2(\alpha-A)+z^2)^2} e^{-z^2/2} dz + \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{2z^2}{(2(\alpha-A)+z^2)} e^{-z^2/2} dz \\ &= -E_A + \frac{2}{2\alpha+1} \end{aligned} \tag{A.7}$$

Now we calculate E_A and integrate by parts:

$$E_A = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)(2z)}{(2(\alpha-A)+z^2)^2} z e^{-z^2/2} dz = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)(1-z^2)}{(2(\alpha-A)+z^2)} e^{-z^2/2} dz.$$

Again using the equality in (A.1), we obtain

$$\begin{aligned} E_A &= \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{2\alpha+1}{(2(\alpha-A)+z^2)} e^{-z^2/2} dz - 1 \\ &= \frac{1}{2(\alpha-A)} \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(2\alpha+1)((\alpha-A)+z^2-z^2)}{(2(\alpha-A)+z^2)} e^{-z^2/2} dz - 1 \\ &= \frac{2\alpha+1-1}{2(\alpha-A)} - 1 = \frac{\alpha}{\alpha-A} - 1. \end{aligned} \tag{A.8}$$

Combining (A.6)–(A.8), we obtain (3.17). □

Lemma A.4. *For all $\alpha > 0$, we have $A'(\alpha) > 0$ and*

$$\frac{2\alpha}{2\alpha+3} < A(\alpha). \tag{A.9}$$

Proof. It suffices to show that the right-hand side of (3.17) is positive for all $\alpha > 0$, which is equivalent to (A.9). We consider the function

$$g(\alpha) := A(\alpha) - \frac{2\alpha}{2\alpha+3}$$

and show that $g(\alpha) > 0$ for all $\alpha > 0$. Assume this is not true. Since $g(\alpha) > 0$ for all sufficiently small $\alpha > 0$ (due to the asymptotics in Lemma A.1) and $\lim_{\alpha\to\infty} g(\alpha) = 0$ (due to the asymptotics in Lemma A.2), this would imply that

$$g'(\alpha) = 0, \quad g(\alpha) \le 0 \quad \text{for some } \alpha > 0. \tag{A.10}$$

Using the fact that $\left(\frac{2\alpha}{2\alpha+3}\right)' > 0$ and applying Lemma A.3, we have

$$g'(\alpha) < A'(\alpha) = \frac{2\alpha(A-1)+3A}{(2\alpha+1)A}.$$

Since $A(\alpha) \leq \frac{2\alpha}{2\alpha+3}$ for α in (A.10), we obtain

$$g'(\alpha) < \frac{1}{(2\alpha+1)A}\left(2\alpha\left(\frac{2\alpha}{2\alpha+3}-1\right)+\frac{6\alpha}{2\alpha+3}\right) = 0,$$

which contradicts (A.10). □

Acknowledgements. Both authors would like to thank the DFG project SFB 910. The research of the first author was also supported by the DFG Heisenberg Programme and by the "RUDN University Program 5-100".

References

[1] M. Abramowitz, I. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series, **55**, 1965.

[2] C. Bishop. Pattern Recognition and Machine Learning. Springer, 2006

[3] Y. Gal. Uncertainty in Deep Learning. PhD thesis, University of Cambridge, 2016.

[4] J. Gerritsma, R. Onnink, and A. Versluis. Geometry, resistance and stability of the delft systematic yacht hull series. In International Shipbuilding Progress, **28** (1981), 276–297.

[5] P. Gurevich, H. Stuke. Learning uncertainty in regression tasks by deep neural networks. arXiv:1707.07287 [stat.ML] (2017).

[6] D. Harrison, D. L. Rubinfeld. Hedonic prices and the demand for clean air, J. Environ. Economics and Management, **5** (1978), 81–102.

[7] J. M. Hernández-Lobato, R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. arXiv:1502.05336 [stat.ML] (2015).

[8] G. Hinton, D. V. Camp. Keeping neural networks simple by minimizing the description length of the weights. In Proceedings of the Sixth Annual Conference on Computational Learning Theory (1993), 5–13.

[9] J. T. G. Hwang, A. A. Ding, Prediction intervals for artificial neural networks, J. Amer. Stat. Assoc, **92**, No. 438 (1997), 748–757.

[10] P. Jylänki, A. Nummenmaa, A. Vehtari. Expectation propagation for neural networks with sparsity-promoting priors. The Journal of Machine Learning Research, **15** (2014), 1849–1901.

[11] H. Kaya, P. Tüfekci , S. F. Gürgen: Local and global learning methods for predicting power of a combined gas and steam turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE (2012), 13–18.

[12] A. Kendall, Y. Gal. What uncertainties do we need in Bayesian deep learning for computer Vision?, 31st Conference on Neural Information Processing Systems (NIPS 2017).

[13] A. Khosravi, S. Nahavandi, D. Creighton, A. Atiya, Comprehensive review of neural network-based prediction intervals and new advances, IEEE Trans. Neural Networks, **22**, No. 9 (2011), 1341–1356.

[14] B. Lakshminarayanan, A. Pritzel, C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.

[15] B. Lakshminarayanan, A. Pritzel, C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[16] M. Lichman. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2013).

[17] A. Lucas. Outlier Robust Unit Root Analysis. PhD Thesis. 1996.

[18] D. MacKay. A practical Bayesian framework for backpropagation networks. Neural computation, **4** (1992), 448–472.

[19] K. Murphy. Machine Learning. A Probabilistic Perspective. MIT Press. Cambridge, 2012.

[20] P. Myshkov, S. Julier. Posterior distribution analysis for Bayesian inference in neural networks. Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.

[21] S. Nadarajah, S. Kotz. Estimation methods for the multivariate t-distribution. Acta Appl Math **102** (2008), 99–118.

[22] R. Neal. Bayesian learning for neural networks. PhD thesis, University of Toronto, 1995.

[23] D. A. Nix and A. S. Weigend, Estimating the mean and variance of the target probability distribution, in Proc. IEEE Int. Conf. Neural Netw., **1**. Orlando, FL, Jun.-Jul. 1994, pp. 55–60.

[24] I. Ortigosa, R. Lopez and J. Garcia. A neural networks approach to residuary resistance of sailing yachts prediction. In Proceedings of the International Conference on Marine Engineering MARINE 2007, 2007.

[25] C. Scheffler. A derivation of the EM updates for finding the maximum likelihood parameter estimates of the Student's t distribution. Working Paper, http://www.inference.org.uk/cs482/publications/scheffler2008derivation.pdf (2008).

[26] J. Soch, C. Allefeld. Kullback–Leibler divergence for the normal-gamma distribution. arXiv:1611.01437 [math.ST] (2016).

[27] D. M. Titterington. Bayesian methods for neural netowrks and related models. Statistical Science, **19**, No. 1 (2004), 128–139.

[28] P. Tüfekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power and Energy Systems, **60** (2014), 126–140.

[29] G. Uludag, A. S. Uyar, K. Senel, H. Dag, Comparison of evolutionary techniques for value-at-risk calculation. In: Giacobini M. (eds) Applications of Evolutionary Computing. EvoWorkshops 2007. Lecture Notes in Computer Science, vol 4448. Springer, Berlin, Heidelberg. 2007.

[30] R. D. De Veaux, J. Schumi, J. Schweinsberg, L. H. Ungar, Prediction intervals for neural networks via nonlinear regression, Technometrics, **40**, no. 4, (1998), 273–282.

[31] C. J. Wild and G. A. F. Seber, Nonlinear Regression. New York, Wiley, 1989.

[32] I-C. Yeh. Modeling of strength of high performance concrete using artificial neural networks, Cement and Concrete Research, **28**, No. 12 (1998), 1797–1808.