Bayesian deep learning

Insights in the Bayesian paradigm for deep learning

Master's Thesis Applied Mathematics

W.R. Schipper

Delft University of Technology

TU Delft

Photo credit: Kitzsteinhorn

Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science

Master's Thesis

Applied Mathematics

Stochastics and Mathematics of Data Science

Bayesian Deep Learning

Insights in the Bayesian paradigm for deep learning

Author:
W.R. Schipper

Supervisor:
Prof. dr. A.W. van der Vaart

Thesis committee member:
Dr. A. Heinlein

Wednesday 30th August, 2023

Abstract

In this thesis, we study a particle method for Bayesian deep learning. In particular, we look at the estimation of the parameters of an ensemble of Bayesian neural networks by means of this particle method, called Stein variational gradient descent (SVGD). This method iteratively updates a collection of parameters and it has the property that its update directions are chosen such that they optimally decrease the Kullback-Leibler divergence. We also study gradient flows of probability measures and show how gradient flows corresponding to functionals on the space of probability measures can induce particle flows. We formulate SVGD as a method in this space. In the regime of infinite particles we show results about convergence of SVGD. An existing convergence result for SVGD can be extended by showing that the probability measures, governing the collection of SVGD particles, are uniformly tight. We give conditions under which this holds.

Preface

Writing a master's thesis in applied mathematics is a far from easy journey. Let me compare it to ascending a snowy mountain in the Alps. It is a long and lonely trip without an easy path that guides you to the top. In fact, you have to find your own path and explore it: see where it takes you. During this trip, you will experience many shivering moments, poor visibility and fatigue. There are only three ingredients that can help you to reach the top: curiosity, determination and stamina. These factors are not purely intrinsic to me, no. To be able to get to the top, I need my family and friends. In particular, I want to thank my parents, Carolien and Joep, for always being there for me, even though I was not always there for them. Furthermore, I would like to thank my sister, Willemijn, for her sweetness and I want to express my gratitude towards my grandmother, Ellen. I do not know anyone else who was so interested in this journey and my experiences throughout. I also want to thank all my friends for making this journey possible, even though this journey also meant that I could not always attend social events, especially in Groningen. It was not easy. For giving me valuable feedback I want to thank Chris van Vliet. Every successful journey has a guiding force and in this case I would like to warmly thank the supervisor of this trip: prof. dr. A.W. van der Vaart. Without his patience, sharp and clear view on the way up, the top would not have been reached.

As time progressed, the journey progressed as well. At this moment in time, I have almost reached the top. At that point, I will finally be able to enjoy the magnificent view from the top of the mountain and see where this journey led me: looking back is sometimes good to see how far you have come. However, the best is yet to come: skiing down the mountain, in a rhythmic flow, determined by the steepness and the curvature of the mountain, as a particle that is smoothly rolling down. A feeling of enormous joy overtakes your thoughts, your sense of being. In no time you will be at the foot of the mountain again: the starting point of the journey. This is what mathematics is all about. A tough and long way up, but then, the hard work pays off and you get a pulse of pure joy: understanding some minuscule part of the mathematical world. I hope this thesis takes you on this journey.

Wieger Schipper

Delft, August 2023

Contents

Notation

Below we will introduce the notation used in this thesis.

Notation	Description
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
P	Probability distribution
$\mathrm{E}, \mathrm{E}_P, \mathrm{E}_{X\sim P}$	Expectation, expectation with explicit dependence on the probability measure P and expectation for the random variable X with distribution P
$\mathcal{D}$	Dataset
$\mathcal{M}$	Model set
p	Number of parameters
p_y	Dimension of output space $\mathcal{Y}$
N, n	Number of datapoints, i.e. sample size
y	Output
x	Input or realisation of random variable X
θ	Parameter
Θ	Parameter space
φ	Density for a normal distribution
M	Number of ensemble members
K	Number of classes in classification problems
$\mathcal{L}$	Loss function
$x \mapsto \eta_\theta(x)$	Regression function
$L(\theta : \mathcal{D})$	Log likelihood for parameter θ and data $\mathcal{D}$
$\|\|\cdot\||, \|\|\cdot\||_2$	General norm and the Euclidean (L_2) norm
$\hat{\theta}^{(m)}$	Estimates of the M parameters in an ensemble, indexed by m
ϵ	Step size
X	Random variable on $\mathcal{X}$
$k(\cdot, \cdot)$	Positive definite kernel function
$\mathcal{H}, \mathcal{H}^d$	Reproducing kernel Hilbert space (RKHS) and a d-dimensional RKHS
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	Inner product with respect to the space $\mathcal{H}$
$\boldsymbol{f}$	Vector-valued function
∇f	Gradient of the real-valued function f
$\nabla \boldsymbol{f}$	Gradient of the vector-valued function $\boldsymbol{f}$, also called the Jacobian
$\boldsymbol{s}_p$	Score function of density p
$\mathcal{A}_p$	Stein operator with respect to p
$\mathcal{Q}$	Set of approximating densities
$p(\cdot)$	Posterior density or target density
$q(\cdot)$	Approximating density
$\mathcal{T}$	Set consisting of smooth transforms
$\mathcal{P}_2(\mathcal{X})$	Wasserstein space on $\mathcal{X}$
$W_2(\nu, \mu)$	Wasserstein-2 distance between μ and ν
$\Rightarrow$	Weak convergence of a measure
$f, \phi. \boldsymbol{f}, \boldsymbol{\phi}$	Real-valued and vector-valued (test) functions
$\|\|\cdot\||^*$	Dual norm
$C_c^\infty(\mathcal{X})$	Space of compactly supported and infinitely many times continuously differentiable functions on $\mathcal{X}$
$(a)^+$	Positive part of $a \in \mathbb{R}$

1 Introduction

In this section an introduction is given to the topic of Bayesian deep learning from a mathematical perspective. A broader view on Bayesian learning is used to motivate why a Bayesian approach can be beneficial in a deep learning setting.

1.1 A motivation for Bayesian deep learning

The introduction that follows is mainly based on Andrew Gordon Wilson 2020 and Andrew G Wilson and Izmailov 2020 and we will also use their notation. In both papers an insightful connection between Bayesian learning and deep learning is made, but let us start this discussion with an interesting motivating problem for Bayesian learning in general. Consider Figure 1, which shows data of the number of monthly airline passengers over the years. See Appendix B.1 for an interesting plot to see what happens when you connect the datapoints.

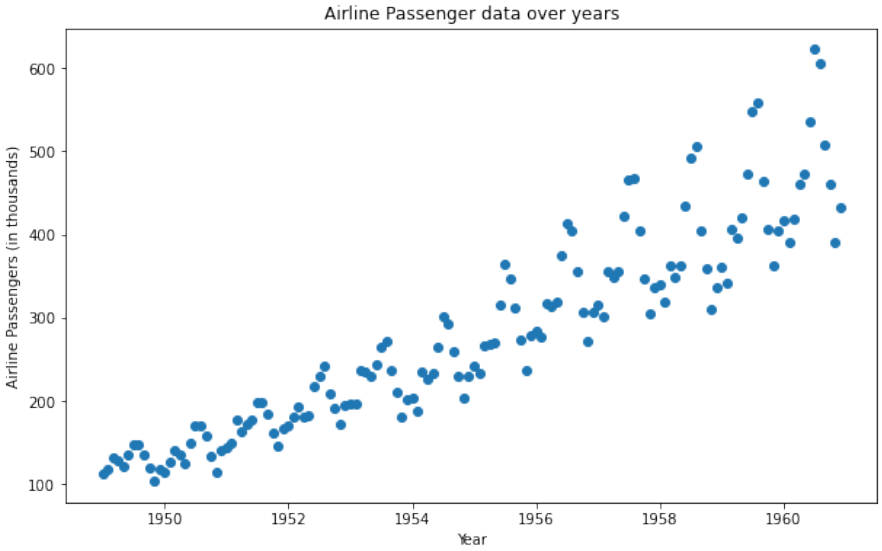

Figure 1: Data for the number of airline passengers per month in the displayed years. Picture inspired by Andrew G Wilson and Izmailov 2020.

The goal is to find the best fitting curve for this data set, i.e. to find a description of the number of airline passengers over time. Consider three possible functions f_k for $k = 1, 2, 3$ for the input x:

$$x \mapsto f_1(x) = \theta_0 + \theta_1 x, \quad x \mapsto f_2(x) = \sum_{j=0}^{3} \theta_j x^j, \quad x \mapsto f_3(x) = \sum_{j=0}^{10^4} \theta_j x^j. \tag{1}$$

The $\{\theta_j\}$ are the function parameters that need to be estimated based on the data. A good question to pose is the following: *what is the correct curve for the data in Figure 1 that represents our beliefs about the truth in the best possible way?* Many people would argue for models 1 and 2 being the correct specifications for the data at hand, due to a more parsimonious explanation of the data, compared to model 3. This is known as Occam's razor. It is a well-known principle for

aiming towards more parsimonious (scientific) explanations and theories, see e.g. C. Rasmussen and Ghahramani 2000. According to Occam's razor, one should look for the most 'simple' hypothesis that explains the given data or problem at hand. However, in this setting it is far from certain that the true model is actually given by f_1, f_2 or f_3. We should observe that f_1 and f_2 are special cases of the more general model f_3. Hence, f_3 is able to describe more phenomena as it has the ability to form a more flexible curve and hence is a 'richer' hypothesis.

To make an analogy with modern deep learning, we can view f_1 and f_2 as relatively simple models and f_3 as a very sophisticated deep learning model, e.g. a deep neural network with an incredibly large number of parameters. In fact, a single hidden layer neural network (NN) is shown to be a universal approximator, see e.g. (Hornik et al. 1989). Informally speaking, this means that it forms a very powerful and sophisticated function approximator. Another type of model with universal approximation guarantees are Gaussian processes (GPs), which are non-parametric models involving infinitely many parameters, see e.g. Williams and C. E. Rasmussen 2006. It is often the case that these GPs output simple predictive distributions. Hence, a large (or small) number of parameters does not necessarily give information about the generalisation behaviour of the model (Andrew G Wilson and Izmailov 2020).

In Andrew G Wilson and Izmailov 2020 it is argued from a probabilistic view that generalization depends on two concepts: the *support* of a model and the *inductive biases* of a model. Let us illustrate and clarify these two concepts by means of an illustration in Figure 2. In Figure 2(a) we have on the horizontal axis an abstraction of image datasets in order of increasing structure in the dataset. For example, corrupted CIFAR-10, an image dataset with completely unstructured random noise pixels (Zhang et al. 2021), has less structure than MNIST (LeCun et al. 1998), which is an image dataset of handwritten digits from zero to nine. In other words, unstructured datasets are more towards the left on the horizontal axis, whereas more structured datasets are on the right on the horizontal axis. On the vertical axis the Bayesian evidence $p(\mathcal{D}|\mathcal{M})$ is depicted, i.e. the probability of a certain dataset $\mathcal{D}$, given some model class $\mathcal{M}$ (i.e. a collection of models). It can be calculated as follows $p(\mathcal{D}|\mathcal{M}) = \int_\Theta p(\mathcal{D}|\mathcal{M}, \theta)p(\theta)\mathrm{d}\theta$, with the first term in the integrand, $p(\mathcal{D}|\mathcal{M}, \theta)$ being equal to the likelihood when the model with parameters θ is used. The second term, $p(\theta)$ is the prior over the model parameters θ. These parameters θ are coming from a parameter space Θ. Hence, $p(\mathcal{D}|\mathcal{M})$ should be interpreted as the weighted probability of a model from $\mathcal{M}$ with parameters θ, weighted by the prior $p(\theta)$ for that specific choice of θ. The average is then taken over all possible parameters $\theta \in \Theta$. The support can now be defined as those datasets $\mathcal{D}$ for which we have $p(\mathcal{D}|\mathcal{M}) > 0$. It is a property of the model class $\mathcal{M}$ and we can write the support more formally as $\mathrm{supp}(\mathcal{M}) = \{\mathcal{D} \mid p(\mathcal{D}|\mathcal{M}) > 0\}$. The inductive biases can then be viewed as the distribution over datasets $\mathcal{D}$, given a certain model class $\mathcal{M}$. This distribution is specified by $p(\mathcal{D}|\mathcal{M})$: the Bayesian evidence. It characterizes which datasets $\mathcal{D}$ are a-priori more likely for a specific model class $\mathcal{M}$. In this way, the inductive bias is depending on the model class $\mathcal{M}$, but also the underlying data model to calculate the likelihood.

Ideally, a model has a very big support, so that any possible hypothesis can be captured. For example, the hypothesis that generates the pure noise CIFAR-10 dataset (Zhang et al. 2021) should not be ruled out. A model should not only have big support, but also inductive biases, meaning that the model should have certain a-priori 'preferences' for certain hypotheses. So, given a specific problem, the model should favour certain hypotheses for that problem, because otherwise it does not converge to a particular hypothesis. In other words, inductive bias is the bias or tendency of a model to favour some hypotheses over others in order to be able to perform inductive inference. As an example, consider the problem class of images. Our model needs a preference for hypotheses that have certain statistical properties that are good descriptions of

images, i.e. the model should have a convolutional structure (Andrew G Wilson and Izmailov 2020).

Let us zoom in on Figure 2(a) to explain this better. The purple curve depicts a model consisting of linear models, $x \mapsto f(x) = \theta_0 + \theta_1 x$, together with a prior $p(\theta_0, \theta_1)$ over the parameters θ_0 and θ_1. In turn, this parameter prior induces a prior over functions, as parameters θ_0, θ_1 are sampled from the prior and these parameters induce functions $x \mapsto f(x) = \theta_0 + \theta_1 x$. Observe that the model consisting of these linear functions has a small support, as higher order functions of x cannot be represented by it. The marginal likelihood (evidence) $p(\mathcal{D}|\mathcal{M})$ has to be normalized over $\mathcal{D}$ and hence all probability mass is given to the datasets $\mathcal{D}$ that can be formed by linear functions from $\mathcal{M}$. This is a very limited amount of datasets and hence the inductive bias is quite narrow, as is depicted in the figure. The pink curve represents a large and fully-connected multi-layer perceptron (MLP). This model class is very flexible and can represent many different datasets, but its structure is not particularly compelling for any specific type of image dataset and hence its inductive bias is very broad. The green curve, representing a convolutional neural network (CNN), is a flexible model class and hence has broad support. However, it has a specific inductive bias for image recognition. In other words, a CNN can represent many hypotheses (broad support), but due to its very specific model characteristics (its convolutional layers in the NN) it can model image datasets particularly well. This is the reason that CNNs have a good inductive bias for image problems.

Let us now focus on Figure 2(b),(c) and (d). For models with a large prior hypothesis space, i.e. a large support, it is possible that the posterior contracts around the true hypothesis for a problem. However, if a simple model with a smaller prior hypothesis space is used, then posterior contraction cannot take place around the true solution, as it is simply not in the support. In (d), the model under consideration has a large prior hypothesis space, but posterior contraction is lacking as the model under consideration has weak inductive biases for the given problem. That means, the model does not have strong preferences for hypotheses and its Bayesian evidence $p(\mathcal{D}|\mathcal{M})$ is too evenly distributed on the large support.

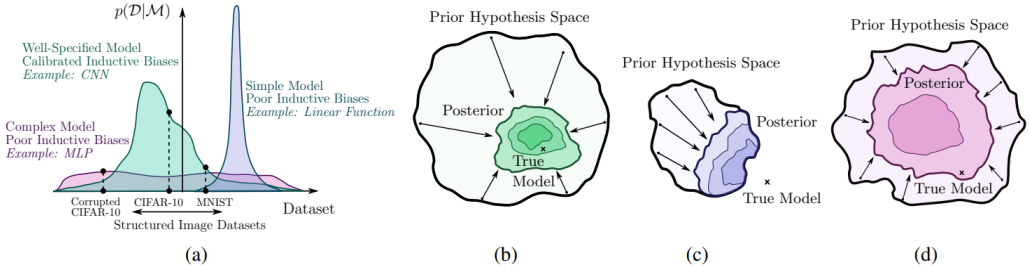

Figure 2: (a) Illustration of inductive biases for complex, simple and well-specified models for differently structured image datasets. The performance of a model always depends on the class of problems under consideration and the inductive biases for each model. For the problem type of structured image datasets, we see that CNNs have well-calibrated inductive biases, whereas a fully connected MLP does not have this well-calibrated inductive bias for these type of problems. Simple linear models also do not have a strong inductive bias for this type of problems. (b) Graphical interpretation of posterior contraction. The large prior hypothesis space is depicted in light-green and the dark green region represents the posterior hypothesis space, which contains the true model. (c) For a misspecified prior hypothesis space, the true model is not in reach and hence the posterior cannot capture the true model. (d) For a large prior hypothesis space, the posterior can capture the true model, but its contraction towards the true model is not very efficient if there are no suitable inductive biases. Picture from Andrew G Wilson and Izmailov 2020.

To come back to the example at the beginning of the section in Figure 1 and the model choice problem, we can now view it from a different perspective. The higher-order polynomial f_3 offers a bigger support, but the choice of prior is very important to model inductive biases. This is still a subjective choice and captures the way in which the modeller thinks about a-priori likely hypotheses. However, what is now more clear is that the choice of model depends on two things: the support and the inductive bias.

1.2 Types of uncertainty

A distinguishing approach of Bayesian inference is that a solution is marginalized over all parameters, weighted by the posterior distribution. This is in contrast to non-Bayesian inference, where a model is formed on the basis of optimization of parameters and not 'weighted'. This yields a full bet on one single hypothesis and no probabilistically weighted average of hypotheses. A neural network is in many cases underspecified by the data, i.e. the number of parameters p is (much) larger than the amount of observations n. This is also called the high-dimensional setting. In this setting it is often the case for NNs that completely different parameter values give rise to different, but also well-performing final networks. In this setting, Bayesian marginalization can make a lot of difference, as it can combine these models in a principled Bayesian way.

In most cases, the predictive density is of interest, as this gives the probability of a new output y, given an input x and the available data $\mathcal{D}$. The output y can e.g. be a class label or a regression value. The input x can for example be an image, the height of a person or the price of a stock. The weights/parameters of the model $f(x;\theta)$ are denoted by θ. In a Bayesian way this predictive density is calculated as follows:

$$p(y|x,\mathcal{D}) = \int_{\Theta} p(y|x,\theta)p(\theta|\mathcal{D})\mathrm{d}\theta, \tag{2}$$

with $p(y|x,\theta)$ being the likelihood (not depending on $\mathcal{D}$) and $p(\theta|\mathcal{D})$ the posterior (not depending on x). In this formulation it is very clear that the Bayesian predictive density is in fact a Bayesian model average (BMA). Every possible setting of parameters θ is used and weighted by means of the posterior $p(\theta|\mathcal{D})$. The parameters θ are marginalized out and thus the predictive density no longer depends on a parameter. The posterior density $p(\theta|\mathcal{D})$ itself is calculated as follows:

$$p(\theta|\mathcal{D}) = \frac{p(\theta,\mathcal{D})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\Theta} p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta} \propto p(\mathcal{D}|\theta)p(\theta). \tag{3}$$

The term $\int_{\Theta} p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta$ is often called the normalization constant, as it does not depend on θ and acts as a normalization term. Observe that this normalization term is, generally speaking, hard to calculate, as it involves an integral over the complete high-dimensional parameter space Θ.

On a more philosophical note, this BMA also represents epistemic uncertainty. Epistemic uncertainty is the uncertainty related to the choice of parameters from the model. It is hard to find exactly which model in $\mathcal{M}$ is true, given the data. Epistemic uncertainty is also called model uncertainty, in contrast to aleatoric uncertainty. Aleatoric uncertainty is the uncertainty inherent in the measurements itself. Consider for example the linear regression model $y = \theta_0 + \theta_1 x + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0,\sigma^2)$. The term ε is the error term in the true linear regression model and this is the aleatoric uncertainty that has nothing to do with the model, but is inherent to the measurement y. This is boldly stated, because we assume here to be a true model, parameterized by θ_0 and θ_1 and for these true parameters there is an aleatoric error term ε involved. Hence, the aleatoric error term also assumes some true underlying model, which can be confused with epistemic uncertainty, as it is hard to infer from measurements y where the uncertainty comes from. Is it the case that we cannot infer the true parameters, based on limited data, and hence have epistemic uncertainty or is it the (stochastic) uncertainty in $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ that we observe? In Figure 3 a predictive model is shown, together with the related epistemic and aleatoric uncertainty. A true function is shown, together with its aleatoric uncertainty that is due to a stochastic error term ε that yields the aleatoric uncertainty. This true function, together with its error term also generated the data points. On the basis of these limited data points, the predictive model has to form a prediction of the true function, which is shown in orange. There is uncertainty involved in these predictions of the underlying model and that is shown in shaded orange.

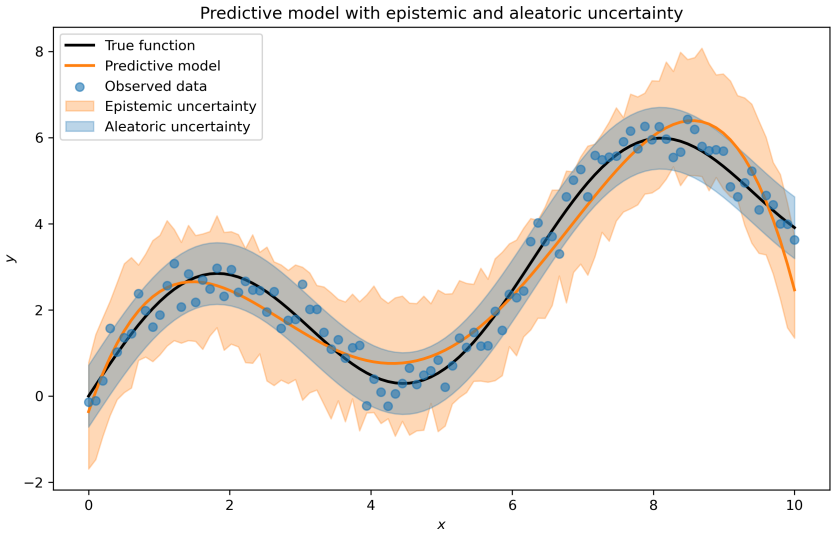

Figure 3: Illustration of a predictive model and its related epistemic and aleatoric uncertainty. The true data generating function is shown (in black) with its related stochastic aleatoric uncertainty in the blue shaded region. A predicted model uses the observed data to infer this true function. However, on the basis of limited data, uncertainty in the inference for the true data generating model is inherited in the form of epistemic uncertainty, depicted in the orange shaded region.

From this perspective there is a big difference between epistemic and aleatoric uncertainty, but it depends on the perspective. From a data perspective, i.e. when you only have the data y available, then it is hard to make a clear distinction between the two. In Figure 4 the dependence of the epistemic and aleatoric uncertainty on the sample size n is depicted. With increasing sample sizes, more complex models come into play and the epistemic uncertainty related to these models also changes accordingly. The aleatoric uncertainty is not affected by this, as it is not related to sample size. This figure illustrates that epistemic uncertainty is intertwined with sample size and the corresponding model complexity belonging to larger sample sizes. Furthermore, the aleatoric uncertainty, that cannot be inferred from simply looking at y, is substantially different from the epistemic uncertainty. The reason we made this distinction is that taking a Bayesian perspective is quite suitable for representing epistemic uncertainty, as the BMA with its weighted posterior formalizes this epistemic uncertainty, given the available data.

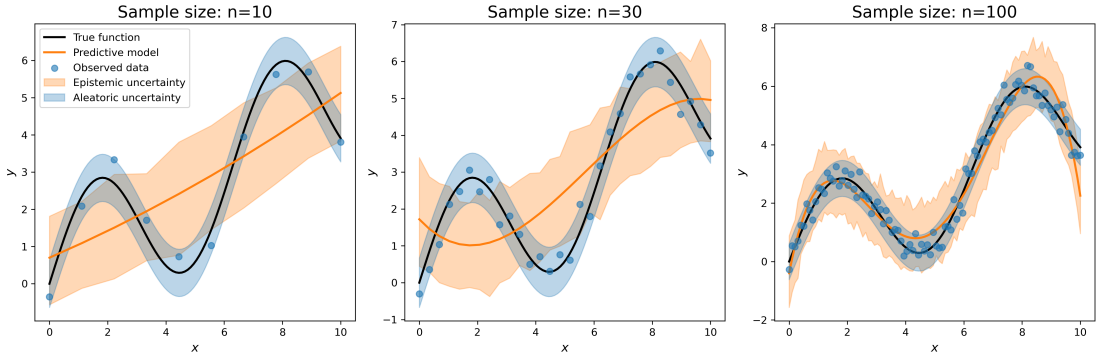

Figure 4: Illustration of the true data generating function with related stochastic aleatoric uncertainty in the blue shaded region. A predictive model is used on the observed data to infer this true function. However, on the basis of limited data, uncertainty in the inference for the true data generating model is inherited in the form of epistemic uncertainty, depicted in the orange shaded region.

1.3 Towards deep learning

A very well-known way of estimating the parameters in classical training procedures is finding a regularized maximum likelihood (ML) solution:

$$\hat{\theta} = \arg\max_{\theta}\{\log p(\theta|\mathcal{D})\} = \arg\max_{\theta}\{\log p(\mathcal{D}|\theta) + \log p(\theta) + C_{\mathcal{D}}\}, \text{ for some constant } C_{\mathcal{D}} \in \mathbb{R}, \quad (4)$$

where we have used that $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$ and that the constant $C_{\mathcal{D}}$ can depend on the data, expressed by the subscript. In a Bayesian setting this procedure is called maximum a-posteriori (MAP) estimation. This naming is well-chosen as this procedure comes down to finding the parameter θ that maximizes the posterior. The log likelihood $\log p(\mathcal{D}|\theta)$ is created by means of linking the output function of our model, denoted $x \mapsto f(x;\theta)$, to the dataset $\mathcal{D}$. To be able to make probabilistic statements, it is needed to assume an underlying true statistical model. Let us give an example to clarify this.

Example 1.1. Consider a regression with Gaussian noise in which we want to model (the mean of) the regression function. The model is $y_j = f(x_j;\theta) + \varepsilon_j$ for $j = 1, \ldots, n$ with $\varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. The likelihood is then given as $p(\mathcal{D}|\theta) = \prod_{j=1}^{n} p(y_j|x_j,\theta) = \prod_{j=1}^{n} \varphi(y_j; f(x_j;\theta), \sigma^2)$, with φ denoting the density of a normally distributed random variable. In this case, the log-likelihood can be seen as a (scaled) mean squared error (MSE) loss: $\log p(\mathcal{D}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - f(x_j;\theta))^2$.

In equation (4) we can view the prior $p(\theta)$ as a regularizer on the parameters θ. Compare it for example to LASSO or Ridge regression (see e.g. Hastie et al. 2009 for more details on these regression models). The prior determines how much of a 'penalty' every value of θ gets in the sense that a low prior value for some θ (more penalty) results in less contribution to finding $\hat{\theta}$. A-priori unlikely θ are more unlikely to become $\hat{\theta}$, so in that sense the prior acts as a regularizer. Observe that MAP estimation is not at all a Bayesian method, as we end up with one single

estimate $\hat{\theta}$, instead of a distribution over parameters. With MAP estimation the final hypothesis is simply $f(x;\hat{\theta})$, with $\hat{\theta}$ given by equation (4).

The major conceptual difference between classical and Bayesian approaches is that the latter generates a probability distribution over the parameters, whereas the classical approach only generates point estimates. However, the difference between classical and Bayesian approaches for estimating the parameters also depends on the shape of the posterior, as a more narrowly peaked posterior more closely resembles MAP estimation due to the fact that it is closer to a point mass. In contrast, a more evenly distributed posterior is 'more' Bayesian in the sense that it really distributes its mass for different values of θ. This diffuse posterior is then definitely not properly captured by a single MAP estimate, which is simply a point mass. This makes the difference between the classical and Bayesian method larger in practice.

In the deep NN setting it is quite often the case that the NN is underspecified by the available data, i.e. the network has too many parameters and too little data: the high-dimensional setting. This results in a non-sharply peaked likelihood $p(\mathcal{D}|\theta)$. Furthermore, in this high-dimensional setting it is also often the case that different settings of the parameters yield different, but well-performing networks on the data. This has for instance been observed in Garipov et al. 2018. This empirical observation about different parameter settings yielding different, but high-performing neural networks, can be a reason to combine different high-performing neural networks in an ensemble. In fact, a Bayesian model average naturally gives rise to an ensemble, weighted by the posterior.

In Lakshminarayanan et al. 2017, an ensemble of deep neural networks is shown to perform very well and it is a highly cited paper (more than 4000 citations). The ensemble of deep neural networks is called a deep ensemble. A deep ensemble is formed by estimation of a neural network model multiple times, where each initial set of parameters is different. Then, after training it is ideally the case that these different parameter initialisations end up in different (local) minima of the loss function. This procedure can be seen as an approximate Bayesian model average as follows. If the prior on the parameters is taken to be a specific normal distribution, then the estimates of θ given by the minimization of the loss function used to train the neural network model in Lakshminarayanan et al. 2017 coincide with (local) MAP estimates when this specific prior is used for θ. Let us consider the case that we take M neural networks in our ensemble. Each of these M models will be trained with the loss function, but in this specific case with a normal prior on θ they form MAP estimates. Hence we end up with M estimates of the parameters of the NN: $\hat{\theta}_1, \ldots, \hat{\theta}_M$. So instead of one single point mass estimate $\hat{\theta}$ in classical training, we now have M estimates. These estimates form a collection of point masses, with each of them equal to a MAP estimate. These estimates can be used to form an approximation of the posterior. The goal is then to approximate the BMA of equation (2), where we now have M MAP estimates forming an approximation of the full posterior. In this way, an ensemble of neural networks can be seen as an approximate Bayesian model average. In Section 2 more details are given.

The estimates $\hat{\theta}_1, \ldots, \hat{\theta}_M$ induce functions $f(\cdot\,;\hat{\theta}_1), \ldots, f(\cdot\,;\hat{\theta}_M)$. It is important that these functions are diverse in their predictions. This diversity gives a better approximation to the BMA, as in this integral the terms $p(y|x,\theta)$ are being added. Let us explain this. Consider two parameter settings θ_1 and θ_2 that are non-equal (and both have positive posterior probability). In the BMA the terms $p(y|x,\theta_1)$ and $p(y|x,\theta_2)$ are added, but if the parameters θ_1 and θ_2 induce two functions $f(\cdot\,;\theta_1)$ and $f(\cdot\,;\theta_2)$ that yield (very) similar predictions, then their induced likelihoods $p(y|x,\theta)$ will also be very similar for both parameters. Hence, in the BMA these two different parameters yield two very similar likelihood terms $p(y|x,\theta)$ that does not result in a diverse ensemble of likelihood terms.

In this way, we can view deep ensembles as an approximate BMA and hereby representing multiple basins of attraction in the posterior. It is the case that most Bayesian methods in deep learning focus on the approximation of the posterior in the neighborhood of one single basin of attraction instead of multiple basins. In Figure 5 an illustration is given to visualize this. In this figure the difference between deep ensembles, variational inference (VI) and Multi-SWAG is illustrated. VI is a standard variational single basin approach and Multi-SWAG is a mixture of Gaussian approximations of the posterior, where each Gaussian is centred on a different basin of the posterior. In a sense, it is a combination of VI and deep ensembles. This method is proposed in Andrew G Wilson and Izmailov 2020.

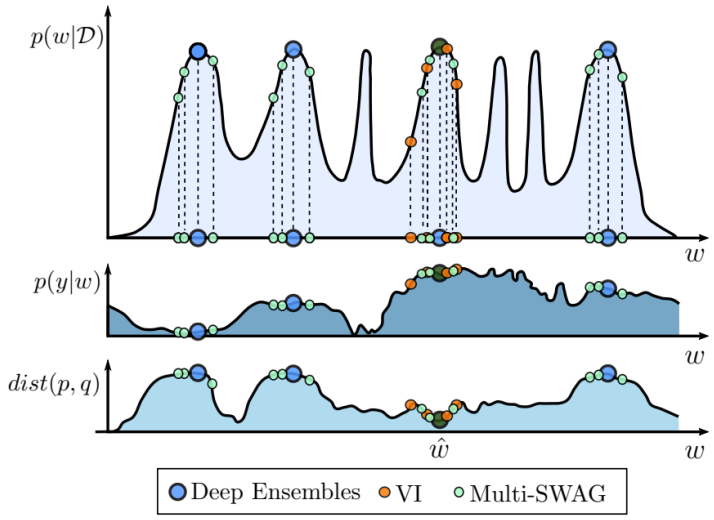

Figure 5: In this illustration the parameter θ is denoted as w. **Top:** approximation of the posterior $p(\theta|\mathcal{D})$ by deep ensembles, VI and Multi-Swag. Observe that deep ensembles and Multi-SWAG use different basins, whereas VI is limited to one basin and its neighborhood. **Middle:** Given a fixed x, the likelihood $p(y|x,\theta)$ is displayed as a function of θ for the three methods. Observe quite significant changes between the basins. **Bottom:** The decrease in some distance metric d between the true density p and the approximating density q is depicted as a function of representing the posterior by an additional parameter θ, while assuming that the mode (in green on the top picture) is sampled. Picture and caption (adapted) from Andrew G Wilson and Izmailov 2020.

The overall interpretation of Figure 5 is that it is beneficial to explore multiple basins for the approximation of the posterior, as it gives a more faithful approximation of the posterior. The terms $p(y|x,\theta)$ vary quite a lot between different basins and the variety of the terms $p(y|x,\theta)$ is needed to approximate the BMA in equation (2) in a good way. The last panel shows that a big decrease in $dist(p,q)$ can be made by not only using the mode as an approximation for the posterior, but also using a second parameter θ. This decrease is especially large when sampling from other basins (areas in which $p(\theta|\mathcal{D})$ is large, see e.g. the places where the blue and green dots are positioned).

2 Deep ensembles

This section gives the necessary background on deep ensembles and motivates a different perspective on these deep ensembles. In particular, the use of Stein variational gradient descent (SVGD) for training deep ensembles is motivated.

2.1 Overview and training of deep ensembles

Deep ensembles are introduced in Lakshminarayanan et al. 2017. Neural networks have shown good performance in a wide range of tasks. However, uncertainty quantification for neural networks is a challenge, as they are black box predictors. In this paper, an alternative to Bayesian neural networks (BNNs) is proposed. Currently, BNNs are a state-of-the-art modelling approach for uncertainty quantification. In a series of experiments, the authors demonstrate that the proposed deep ensembles produce well-calibrated uncertainty estimates which are as good as, or better than approximate BNNs. The aim of this section is to introduce deep ensembles. To this end, we will closely follow the original paper.

It is assumed that a training set $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \dots, n\}$ of n i.i.d. datapoints is available with $x \in \mathbb{R}^d$ being the d-dimensional features and y being the labels. In classification problems it is assumed that $y \in \{1, \dots, K\}$, i.e. y is in one of the K classes. In regression problems, no restrictions are imposed on y, i.e. $y \in \mathbb{R}^{p_y}$. We mostly consider the case $y \in \mathbb{R}$, i.e. $p_y = 1$ for a simple regression setting. Given the inputs x, a neural network is used to model a probabilistic predictive density function $p_\theta(y|x)$ for the labels y, given an input vector x. The subscript θ denotes the parameters of the NN. Let M denote the number of neural networks collected in the ensemble. The collection of parameters in the ensemble is denoted as $\{\theta_m\}_{m=1}^M$.

Following Hoffmann and Elster 2021, let us give the assumed heteroscedastic regression model:

$$Y|X \sim \mathcal{N}(f(X;\theta), \sigma_\theta^2(X)I), \tag{5}$$

with realisations of X being in $\mathbb{R}^d$ and realisations of Y taking value in $\mathbb{R}^{p_y}$. Throughout this section, $x \mapsto f(x;\theta)$ will be called the regression function. It is often abbreviated as $f_\theta(x) := f(x;\theta)$. To be clear, the available data is $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \dots, n\}$ and this data is sampled from equation (5).

Deep ensembles, as defined in Lakshminarayanan et al. 2017, can be seen as a way to perform approximate Bayesian inference in which the posterior density $\theta \mapsto p(\theta|\mathcal{D})$ of the neural network parameters θ, given the available data $\mathcal{D}$, is approximated by a family of empirical distributions in the following sense:

$$\int_B p(\theta|\mathcal{D})\mathrm{d}\theta \approx \frac{1}{M}\sum_{m=1}^{M} \delta_{\hat{\theta}^{(m)}}(B), \quad \text{for any measurable set } B, \tag{6}$$

with δ_θ the Dirac measure with support $\{\theta\}$. We also need to assume a distribution for the network parameters θ, i.e. we need to assume a prior density $p(\theta)$. This is done as follows:

$$\theta \sim \mathcal{N}(0, \lambda^{-1}I), \tag{7}$$

where λ is the L_2 regularization parameter in the loss function of the neural network in the training phase, see equation (8). For the deep ensembles, we need to train M models, to obtain

(ideally different) estimates of θ, denoted by $\hat{\theta}^{(m)}$ for $m = 1, \ldots, M$. For each of these M models an initial random parameter setting is used, following equation (7). These random initialisations of the parameters are then independently updated as follows. Every neural network parameter initialisation in the ensemble is independently updated (after a random parameter initialisation) by a minimization of the following loss function:

$$\theta \mapsto \mathcal{L}(\theta) = -L(\theta : \mathcal{D}) + \frac{1}{2}\lambda||\theta||_2^2, \tag{8}$$
$$\text{with } L(\theta : \mathcal{D}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{||y_i - f_\theta(x_i)||_2^2}{\sigma_\theta^2(x_i)} + p_y \log(\sigma_\theta^2(x_i))\right),$$

where $||\cdot||_2$ is the standard Euclidean L_2 norm and $L(\theta : \mathcal{D})$ is the log likelihood for the assumed statistical model in equation (5) (up to a constant not depending on θ). The estimates for the parameters are denoted by $\hat{\theta}^{(m)}, m = 1, \ldots, M$, for the M neural networks in our ensemble. It turns out that these obtained parameter estimates are equal to the local MAP estimates of θ when the prior in equation (7) is used, see Appendix C.1 for a derivation.

In the assumption of the posterior approximation in equation (6) we have the average of M distributions at local MAP estimates for θ as the approximation of the true posterior. This yields that the approximation of the posterior predictive density is given as:

$$p(y|x, \mathcal{D}) = \int p(\theta|\mathcal{D})\varphi(y; f_\theta(x), \sigma_\theta^2(x)I)\mathrm{d}\theta \approx \frac{1}{M}\sum_{m=1}^{M}\varphi(y; f_{\hat{\theta}^{(m)}}(x), \sigma_{\hat{\theta}^{(m)}}^2(x)I), \tag{9}$$

with φ denoting the density of a normally distributed (vector) random variable. This average in equation (9) is an average of Gaussian densities. In Lakshminarayanan et al. 2017, the assumption is made that the ensemble prediction is also Gaussian with mean and variance given by the mean and the variance of the expression in equation (9). They use this as approximation, as it is never the case that an average of Gaussian distributions is again Gaussian. The average of Gaussian densities, denoted $\frac{1}{M}\sum_{m=1}^{M}\varphi(f_{\theta_m}(x), \sigma_{\theta_m}^2(x))$ is modelled to have a mean and variance as follows:

$$f_*(x) = \frac{1}{M}\sum_{m=1}^{M} f_{\theta_m}(x), \text{ and } \sigma_*^2(x) = \frac{1}{M}\sum_{m=1}^{M}(\sigma_{\theta_m}^2(x) + f_{\theta_m}^2(x)) - f_*^2(x).$$

By doing this 'extra' approximation, it is easier to calculate quantiles and predictive probabilities.

Even though the approximation of the posterior of true network parameters by means of an average of Dirac distributions is very simple, it appears that deep ensembles outperform many other Bayesian approaches because the deep ensembles are able to explore and find many different modes of the posterior (Hoffmann and Elster 2021). In a way, it is not fair to call deep ensembles in this framework a Bayesian method, as the posterior is not computed by means of a combination of the prior and the likelihood. In fact, a 'posterior' is formed by optimizing parameters with respect to a loss function and combining these parameters by means of an average of Dirac distributions. It is assumed that this modelled average of Dirac distributions approximates the true posterior, when a prior is taken over the parameters according to equation (7).

2.2 Bayesian ensemble training with SVGD

Deep ensembles have been shown to perform well in in terms of predictive performance and uncertainty estimation, see e.g. Lakshminarayanan et al. 2017. They are the main counterpart to Bayesian neural networks when it comes to uncertainty estimation. Deep ensembles average predictive hypotheses, but no guarantees are given for the (functional) diversity between these hypotheses. Furthermore, deep ensembles are not motivated in a Bayesian probabilistic framework, at least not in the original paper Lakshminarayanan et al. 2017. In D'Angelo and Fortuin 2021, it is shown how a repulsive term between the ensemble members can help to generate functional diversity by avoiding that the ensemble members end up with the same parameters in parameter space. This repulsive term is inspired by Stein variational gradient descent (SVGD). The authors also show how this procedure can be seen as a gradient flow of the Kullback-Leibler (KL) divergence in Wasserstein space, a space of distributions. The authors argue that this reformulation of deep ensembles with repulsion is a Bayesian method.

A big problem with NNs, when working in weight space, is that different parameter weights θ_1 and θ_2 for the NN can induce the same NN function. Hence, 'diversity' in parameter space does not mean diversity in function space. A diverse (parameter) space ensemble is therefore not guaranteed to be truly diverse in function space. This is what the authors call non-identifiability of neural networks.

Let $g : (x, \theta) \mapsto f(x; \theta)$ be the mapping that takes a data point $x \in \mathcal{X}$ and a parameter weight vector $\theta \in \mathbb{R}^d$ to the corresponding neural network output $f(x; \theta)$. Let us denote $f_i := f(\cdot; \theta_i)$ as the neural network output with a certain configuration of weights θ_i and a certain input x. Then, for any non-identifiable pair $\theta_i, \theta_j \in \mathbb{R}^d$ and f_i, f_j their induced neural network functions:

$$f_i = f_j \implies \theta_i = \theta_j.$$

Formally, this means that the map g is not injective. In deep learning, the likelihood function $p(y|f(x; \theta))$ is often considered. A Gaussian regression model or a categorical model for classification are examples of such a likelihood function. See e.g. equation (5) for the Gaussian regression model. This likelihood function is parametrised by the output of the NN, denoted as $f(x; \theta)$. An example would be a neural network outputting the predicted mean and the variance of the Gaussian regression model in equation (5). In this sense the NN parametrises the likelihood function. Let us remind ourselves that we have a set of i.i.d. training data which we denote as $\mathcal{D} = \{(x_i, y_i)|i = 1, \ldots, n\}$. For BNNs, the posterior distribution is of interest. More precisely, one is interested in the posterior density:

$$p(\theta|\mathcal{D}) \propto \prod_{i=1}^{n} p(y_i|f(x_i; \theta))p(\theta),$$

with $p(\theta)$ the prior on the NN parameters. As stated before, the quantity of interest in Bayesian inference is the BMA:

$$p(y|x, \mathcal{D}) = \int_{\Theta} p(y|f(x; \theta))p(\theta|\mathcal{D})\mathrm{d}\theta.$$

Ensembles of neural networks are usually trained by means of maximum a posteriori (MAP) estimation. The non-convexity of the MAP estimation (or optimization) problem is used by the

deep ensembles to form M independently trained (and ideally also different) parameter solutions. Consider M parameter weights of NNs in an ensemble, $\{\theta_m\}_{m=1}^M$ with $\theta_m \in \mathbb{R}^d \, \forall m = 1, \ldots, M$. The evolution of the parameters of the ensemble members under the gradient of the log-posterior gives the following update rule at iteration $\ell \in \mathbb{N}$:

$$\begin{aligned} \theta_m^{\ell+1} &\leftarrow \theta_m^\ell + \epsilon_\ell \phi(\theta_m^\ell), \quad \forall m = 1, \ldots, M, \\ \text{with } \phi(\theta_m^\ell) &= \nabla_\theta \log p(\theta|\mathcal{D})|_{\theta=\theta_m^\ell}, \end{aligned}$$

with (small) step size ϵ_ℓ. In this training procedure there is no constraint imposed to make sure that different ensemble members cannot converge to the same mode of the posterior and hence end up with the same parameters. This is problematic, as this means that having more members in the ensemble does not necessarily mean that the ensemble gets more diverse. The only way in which parameters will not coincide relies (exclusively) on:

- Random initialisation
- The noise in the estimation of the gradients
- The number of local optima that are reached during gradient descent.

The aim of the paper D'Angelo and Fortuin 2021 is to overcome these pitfalls and make stronger guarantees to counteract ending up with equal parameter settings for different ensemble members. Inspired by SVGD (Liu and D. Wang 2016), the authors introduce a repulsive component in the training of the ensemble members. This repulsive component is integrated by means of a kernel function that models a repulsive action if two ensemble members are close to each other in weight space. This prevents that the different ensemble members end up with the same parameters. In Section 3 we will study SVGD and show how it can be a potential training solution for deep ensembles.

3 SVGD

In Bayesian inference the posterior is of main interest. However, computing the posterior may be hard and intractable, due to the normalization constant in the posterior density (3). Markov chain Monte Carlo (MCMC) is a popular method to draw samples from this intractable posterior distribution, see e.g. Brooks et al. 2011. A different type of algorithm, which aims to find an approximation of the posterior, is called variational inference (VI). The idea with VI is to approximate the posterior by means of a simpler distribution in some pre-defined distribution class. The best approximating distribution is found by minimizing the KL divergence between the true posterior and the approximating distribution.

VI can be computationally more efficient than MCMC methods and in Liu and D. Wang 2016 a new and general VI algorithm is developed, called Stein variational gradient descent (SVGD). It can be seen as a gradient descent type of algorithm for Bayesian inference. In the algorithm a set of particles is used to approximate the posterior. A form of gradient descent is applied to these particles with the goal to minimize the KL divergence and evolve these particles in such a way that they approximate the posterior distribution.

Let us adopt the preliminaries of Liu and D. Wang 2016 and introduce theirs. We will now not use θ anymore, but X, a random variable or parameter with support $\mathcal{X} \subset \mathbb{R}^d$. The data is given as $\mathcal{D}$ and is a set of i.i.d. observations. The Bayesian prior is denoted as p_0 and the posterior density is given by p. We slightly abuse notation by writing p instead of $p(\cdot|\mathcal{D})$. In what follows, we will continue writing p to denote the posterior density.

We use the convention that $\mathcal{X} = \mathbb{R}^d$, unless specifically stated otherwise. We define $L^2(\mu) := \{\boldsymbol{f} : \mathcal{X} \to \mathcal{X} \mid \int ||\boldsymbol{f}(x)||^2 \mathrm{d}\mu(x) < \infty\}$ and denote by $\langle \cdot, \cdot \rangle_{L^2(\mu)}, || \cdot ||_{L^2(\mu)}$ its inner product and norm, respectively. We let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a general positive definite kernel function. A function $(x, x') \mapsto k(x, x')$ is positive definite if $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0$ for any $x_1, \ldots, x_n \in \mathcal{X}, n \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathbb{R}$. A reproducing kernel Hilbert space (RKHS), denoted $\mathcal{H}$, with respect to the kernel k is the completion (with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined below) of the linear span of kernel functions, i.e. the completion of the set $\{f : x \mapsto f(x) = \sum_{i=1}^n a_i k(x, x_i), a_i \in \mathbb{R}, n \in \mathbb{N}, x_i \in \mathcal{X}\}$. We equip $\mathcal{H}$ with the inner product $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^n a_i b_j k(x_i, x_j)$ for $x \mapsto g(x) = \sum_{j=1}^n b_j k(x, x_j)$. The space $\mathcal{H}^d$ is used to denote the space of vector functions $\boldsymbol{f} = [f_1, \ldots, f_d]^T$ with $f_i \in \mathcal{H} \, \forall i = 1, \ldots, d$. The inner product on $\mathcal{H}^d$ is $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}}$. All vectors are assumed to be column vectors, unless stated otherwise. For general vector-valued functions $\boldsymbol{f}$, i.e. functions $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^{d'}$, the gradient (Jacobian) of $\boldsymbol{f}$, denoted $\nabla \boldsymbol{f}$, is

$$\nabla \boldsymbol{f} = \begin{bmatrix} \dfrac{\partial \boldsymbol{f}}{\partial x_1} \\ \vdots \\ \dfrac{\partial \boldsymbol{f}}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_{d'}}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_1}{\partial x_d} & \cdots & \dfrac{\partial f_{d'}}{\partial x_d} \end{bmatrix}, \tag{10}$$

or in short-hand notation $\nabla \boldsymbol{f} = \left[\frac{\partial f_j}{\partial x_i}\right]_{ij}$ for $i = 1, \ldots, d$ and $j = 1, \ldots, d'$. For a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$, the gradient of f, denoted ∇f, is simply the conventional column vector. Throughout this section we will assume that $\mathcal{X} \subseteq \mathbb{R}^d$.

We will closely follow Liu, Lee, et al. 2016. Let us first introduce Stein's operator and the Stein

class of a density. After that we state Stein's identity. These are the necessary ingredients for a thorough understanding of SVGD, which is the aim of this section.

Definition 3.1 (Definition 2.1 from Liu, Lee, et al. 2016)**.** Let p be a continuously differentiable density with support $\mathcal{X}$. The score function of p is defined as:

$$\boldsymbol{s}_p = \nabla \log p = \frac{\nabla p}{p}.$$

A function $\phi : \mathcal{X} \to \mathbb{R}$ is in the Stein class of p if ϕ is continuously differentiable and satisfies:

$$\int_{\mathcal{X}} \nabla(\phi(x)p(x))\mathrm{d}x = 0. \tag{11}$$

The Stein operator of p is a linear operator acting on functions in the Stein class of p. For scalar-valued functions ϕ, the Stein operator $\mathcal{A}_p$ is defined as:

$$\phi \mapsto \mathcal{A}_p\phi = \boldsymbol{s}_p\phi + \nabla\phi.$$

A vector-valued function $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_{d'}]^T$ is said to be in the Stein class of p if all ϕ_i for $i = 1, \ldots, d'$ are in the Stein class of p. If $\mathcal{A}_p$ is applied to a vector-valued function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^{d'}$, then $\mathcal{A}_p\boldsymbol{\phi}$ results in a $d \times d'$ matrix-valued function $\mathcal{A}_p\boldsymbol{\phi} = \boldsymbol{s}_p\boldsymbol{\phi}^T + \nabla\boldsymbol{\phi}$.

Remark 3.2. Observe that $\boldsymbol{s}_p$ and $\mathcal{A}_p\phi$ are $d \times 1$ vector-valued functions mapping $\mathcal{X}$ to $\mathbb{R}^d$.

Remark 3.3. Let us consider $\mathcal{X}$, a compact subset of $\mathbb{R}^d$ with a piecewise-smooth boundary $\partial\mathcal{X}$. Denote the unit normal boundary vector of $\partial\mathcal{X}$ by $\boldsymbol{n}$. Then, by a consequence of the divergence theorem it is the case that:

$$\int_{\mathcal{X}} \nabla(\phi(x)p(x))\mathrm{d}x = \oint_{\partial\mathcal{X}} \phi(x)p(x)\,\boldsymbol{n}\mathrm{d}S(x),$$

with $\oint_{\partial\mathcal{X}} \mathrm{d}S$ being the surface integral over $\partial\mathcal{X}$. If $\phi(x)p(x) = 0$ for all $x \in \partial\mathcal{X}$, then equation (11) holds. A more general condition would be that $\oint_{\partial\mathcal{X}} \phi(x)p(x)\,\boldsymbol{n}\mathrm{d}S(x)$=0.

Consider now the case that $\mathcal{X} = \mathbb{R}^d$. The divergence theorem needs a compact set to work on. Let us take $B_r = B(0, r)$, i.e. the closed ball in $\mathbb{R}^d$, centred at 0 with radius r. Denote by ∂B_r its boundary, then by the divergence theorem:

$$\int_{B_r} \nabla(\phi(x)p(x))\mathrm{d}x = \oint_{\partial B_r} \phi(x)p(x)\,\boldsymbol{n}\mathrm{d}S(x).$$

To get the domain of integration equal to $\mathcal{X} = \mathbb{R}^d$, we take $\lim_{r\to\infty} B_r$:

$$\lim_{r\to\infty} \int_{B_r} \nabla(\phi(x)p(x))\mathrm{d}x = \lim_{r\to\infty} \oint_{\partial B_r} \phi(x)p(x)\,\boldsymbol{n}\mathrm{d}S(x).$$

Hence, if we have that $\lim_{||x||\to\infty} \phi(x)p(x) = 0$, then we get that $\lim_{r\to\infty} \oint_{\partial B_r} \phi(x)p(x)\,\boldsymbol{n}\mathrm{d}S(x) = 0$, as $||x|| \to \infty$ is a characterization of $\lim_{r\to\infty} \partial B_r$.

We now have the theory ready to introduce Stein's identity.

Lemma 3.4 (Stein's identity, see e.g. Liu, Lee, et al. 2016)**.** *Let p be a continuously differentiable (i.e. smooth) density supported on $\mathcal{X} \subset \mathbb{R}^d$ and let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^{d'}$, $x \mapsto \boldsymbol{\phi}(x) = [\phi_1(x), \dots, \phi_{d'}(x)]^T$ be a smooth vector-valued function. Stein's identity comes down to:*

$$\mathrm{E}_{X\sim p}[\mathcal{A}_p \boldsymbol{\phi}(X)] = 0, \tag{12}$$

for any $\boldsymbol{\phi}$ that is in the Stein class of p.

Proof. Observe that by the product rule for gradients, we have

$$\frac{1}{p(x)}\nabla_x(\boldsymbol{\phi}(x)p(x)) = \frac{1}{p(x)}((\nabla p(x))\boldsymbol{\phi}(x)^T + p(x)\nabla\boldsymbol{\phi}(x)) = \boldsymbol{s}_p(x)\boldsymbol{\phi}(x)^T + \nabla\boldsymbol{\phi}(x). \tag{13}$$

Because $\boldsymbol{\phi}$ is in the Stein class of p, we have $\int_{\mathcal{X}} \nabla(\phi_i(x)p(x))\mathrm{d}x = 0$ for all $i = 1, \dots, d'$. By this assumption, we can make the term $\int_{\mathcal{X}} p(x)(\boldsymbol{s}_p(x)\phi_i(x) + \nabla\phi_i(x))\mathrm{d}x$ equal to zero for all $i = 1, \dots, d'$. Hence, $\mathrm{E}_{X\sim p}[\mathcal{A}_p\phi_i(X)] = 0$ for all $i = 1, \dots, d'$. We have done this procedure for all $i = 1, \dots, d'$ and hence we can conclude that $\mathrm{E}_{X\sim p}[\mathcal{A}_p\boldsymbol{\phi}(X)] = 0$, where this 0 now represents a $d \times d'$ dimensional matrix. □

The interpretation of this lemma is that the Stein operator $\mathcal{A}_p$ acts on a function $\boldsymbol{\phi}$ and yields an expectation of zero under $X \sim p$. For sufficiently regular $\boldsymbol{\phi}$, i.e. the function should be in the Stein class of p, the Stein identity holds. Consider any other smooth density, different from p. Let us call it q and it also has $\mathcal{X}$ as support. Let us look at the expectation of $\mathcal{A}_p\boldsymbol{\phi}(X)$ when $X \sim q$, i.e. we consider $\mathrm{E}_{X\sim q}[\mathcal{A}_p\boldsymbol{\phi}(X)]$. This quantity is not necessary equal to zero for general, unconstrained $\boldsymbol{\phi}$. In fact, $\mathrm{E}_{X\sim q}[\mathcal{A}_p\boldsymbol{\phi}(X)]$ can be used as a defining measure to quantify the discrepancy between p and q. The bigger $|\mathrm{E}_{x\sim q}[\mathcal{A}_p\boldsymbol{\phi}(x)]|$ is, the more dissimilar p and q are. This gives rise to the so called Stein discrepancy, by measuring the largest deviation from zero of Stein's identity for functions $\boldsymbol{\phi}$ in some pre-defined function set $\mathcal{F}$.

Definition 3.5. Consider a function set $\mathcal{F}$ and two smooth densities p and q. Stein discrepancy $\mathcal{S}$ is defined as:

$$\mathcal{S}(q, p) = \max_{\boldsymbol{\phi}\in\mathcal{F}}\{[\mathrm{E}_{X\sim q}\,\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(X))]^2\}. \tag{14}$$

A different discrepancy measure for two smooth densities, which is inspired by Stein's discrepancy, is kernelized Stein discrepancy (KSD). This KSD is the Stein discrepancy with a special choice of $\mathcal{F}$. The choice of $\mathcal{F}$ is of major importance for Stein's discrepancy due to the fact that it determines the discriminative power and computational tractability of it, see e.g. Liu and D. Wang 2016. KSD considers the functions $\boldsymbol{\phi}$ in the unit ball of a reproducing kernel Hilbert space (RKHS). Furthermore, this KSD has a closed form solution. Let us give the formal definition of KSD:

Definition 3.6. Consider two smooth densities p and q. Consider a kernel $(x, x') \mapsto k(x, x')$ of the RKHS $\mathcal{H}$ that is in the Stein class of the density q. The Kernelized Stein discrepancy, denoted S, is defined in the following way:

$$S(q, p) = \max\{[\mathrm{E}_{X\sim q}(\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(X)))]^2 \quad | \quad \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\}. \tag{15}$$

A kernel $(x, x') \mapsto k(x, x')$ is said to be in the Stein class of q if k has continuous second order partial derivatives and both $k(x, \cdot)$ and $k(\cdot, x)$ are in the Stein class of q for any fixed $x \in \mathcal{X}$.

An analytic result, that gives the function $\boldsymbol{\phi} \in \mathcal{H}^d$ with $||\boldsymbol{\phi}||_{\mathcal{H}^d} = 1$ for which we have the exact number $S(q, p)$, is given as follows:

$$\boldsymbol{\phi} = \frac{\boldsymbol{\phi}^*_{q,p}}{||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d}}, \tag{16}$$

$$\text{with } \boldsymbol{\phi}^*_{q,p}(\cdot) = \mathrm{E}_{X\sim q}[\mathcal{A}_p k(X, \cdot)]. \tag{17}$$

This specific choice of $\boldsymbol{\phi}$ yields the value $S(q, p) = ||\boldsymbol{\phi}^*_{q,p}||^2_{\mathcal{H}^d}$. Let us formalize these two results in the following theorem:

Theorem 3.7 (Theorem from Liu, Lee, et al. 2016)**.** *Let $\mathcal{H}$ be the RKHS related to a positive definite kernel $(x, x') \mapsto k(x, x')$ that is in the Stein class of the density q. Denote by $x' \mapsto \boldsymbol{\phi}^*_{q,p}(x') = \mathrm{E}_{X\sim q}[\mathcal{A}_p k(x', X)]$. Then:*

$$S(q, p) = ||\boldsymbol{\phi}^*_{q,p}||^2_{\mathcal{H}^d}, \tag{18}$$

and the maximum in equation (15) is attained at $\boldsymbol{\phi} = \frac{\boldsymbol{\phi}^*_{q,p}}{||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d}}$. Furthermore, it is the case that $\langle \boldsymbol{\phi}, \boldsymbol{\phi}^*_{q,p} \rangle_{\mathcal{H}^d} = \mathrm{E}_q[\mathrm{tr}(\mathcal{A}_p \boldsymbol{\phi})]$ for any $\boldsymbol{\phi} \in \mathcal{H}^d$.

Proof. See Appendix A.1. □

Another favourable property of $S(q, p)$ is that it is equal to zero if and only if $p = q$ if k is suitably chosen, i.e. strictly positive definite in a proper sense. See Appendix D.1 for more details. Because we established the equality $S(q, p) = ||\boldsymbol{\phi}^*_{q,p}||^2_{\mathcal{H}^d}$, we also have that $\boldsymbol{\phi}^*_{q,p}$ is the zero function (and hence has norm zero) if and only if $q = p$ under the same condition on k. Another useful property of KSD is that it only depends on the unknown target density p via the score function $\nabla \log(p)$. By means of this observation, the score function can be calculated without the need to know the normalization constant Z for the posterior density $x \mapsto p(x|\mathcal{D}) = p_0(x)p(\mathcal{D}|x)/Z$, with p_0 denoting the prior. More precisely, we have $\nabla_x \log p(x|\mathcal{D}) = \nabla_x \log(p_0(x)p(\mathcal{D}|x))$, so Z is not needed.

In variational inference, the goal is to approximate a target density p by using a 'simpler' density q^* from a set of densities $\mathcal{Q} = \{q_j\}_{j\in J}$, for J some indexing set. The approximating density $q^* \in \mathcal{Q}$ is found by minimizing KL divergence. Let us write the posterior density as $p = \bar{p}/Z$, with Z the normalization constant. We have:

$$
\begin{aligned}
q^* &= \arg\min_{q\in\mathcal{Q}}\{KL(q||p)\} \\
&= \arg\min_{q\in\mathcal{Q}}\left\{\int_{\mathcal{X}} \log\left(\frac{q(x)}{p(x)}\right) q(x)\mathrm{d}x\right\} \\
&= \arg\min_{q\in\mathcal{Q}}\left\{\int_{\mathcal{X}} \log\left(Z\frac{q(x)}{\bar{p}(x)}\right) q(x)\mathrm{d}x\right\} \\
&= \arg\min_{q\in\mathcal{Q}}\left\{\int_{\mathcal{X}} (\log(q(x)) - \log(\bar{p}(x)) + \log(Z))q(x)\mathrm{d}x\right\} \\
&= \arg\min_{q\in\mathcal{Q}}\{\mathrm{E}_q[\log q] - \mathrm{E}_q[\log \bar{p}] + \log Z\} \\
&= \arg\min_{q\in\mathcal{Q}}\{\mathrm{E}_q[\log q] - \mathrm{E}_q[\log \bar{p}]\}.
\end{aligned} \tag{19}
$$

Observe that the value of $\log Z$ does not influence the choice of q^*, as it does not depend on q. The choice of the set $\mathcal{Q}$ is crucial, as it determines how close q^* can be to p and how easily computable this optimization problem above is.

In Liu and D. Wang 2016, the authors focus on those sets $\mathcal{Q}$ that are composed of densities which are smooth transforms of a certain reference density, i.e. $\mathcal{Q} = \{q_{[\boldsymbol{T}]} \,|\, \boldsymbol{T} \in \mathcal{T}\}$, with $\mathcal{T}$ a set of smooth transforms. So $\mathcal{Q}$ is the set consisting of the densities $q_{[\boldsymbol{T}]}$, where $q_{[\boldsymbol{T}]}$ is the density of $Z = \boldsymbol{T}(X)$ when X has a reference density q_0. The functions $\boldsymbol{T} : \mathcal{X} \to \mathcal{X}$ are continuously differentiable and bijective transforms and come from the set $\mathcal{T}$. The density of Z is given as follows:

$$
z \mapsto q_{[\boldsymbol{T}]}(z) = q_0(\boldsymbol{T}^{-1}(z)) \cdot |\det(\nabla_z \boldsymbol{T}^{-1}(z))|, \tag{20}
$$

for $\boldsymbol{T}^{-1}$ being the inverse of $\boldsymbol{T}$ and $\nabla\boldsymbol{T}^{-1}$ the Jacobian of $\boldsymbol{T}^{-1}$, following the notational convention of equation (10). This definition of the density is only valid if the Jacobian $\nabla\boldsymbol{T}^{-1}$ is nonsingular on its domain. A reason for choosing these specific type of transformed distributions is that they are computationally tractable, meaning that the expectation of functions of the random variables Z with density $q_{[\boldsymbol{T}]}$ can be easily calculated. Let us clarify this. Assume that we have realisations of the random variable Z: $\{z_i\}_{i=1}^n$ for $z_i = \boldsymbol{T}(x_i)$ and the x_i are realisations of a random variable X that has a density q_0. If an expectation with respect to Z has to be evaluated, then an empirical average over $\{z_i\}$ yields an approximation of the expectation we are interested in, i.e. we approximate $\mathrm{E}_{Z\sim q_{[\boldsymbol{T}]}}[h(Z)]$ by $\frac{1}{n}\sum_{i=1}^n h(z_i) = \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{T}(x_i))$, for some function h.

At this point the choice of transform $\boldsymbol{T} \in \mathcal{T}$ is not restricted and for computational reasons it is necessary to make restrictions on $\boldsymbol{T}$. In Liu and D. Wang 2016 a method is proposed that iteratively computes transforms in a way that resembles steepest (gradient) descent in a RKHS. To be able to explain this procedure in detail, some more theory is needed. We will show how the Stein operator, as in Definition 3.1 can be seen as a derivative of the KL divergence.

The goal is to minimize the KL divergence in equation (19), which enables us to compute q^*. In order to achieve that goal, let us consider a very specific form of transform which is obtained by a small perturbation of the identity transform, i.e. $x \mapsto \boldsymbol{T}(x) = x + \epsilon\boldsymbol{\phi}(x)$ for $\boldsymbol{\phi}$ a smooth function that gives the direction of the perturbation, given some input x. The magnitude of the perturbation is given by $\epsilon \in \mathbb{R}$.

In the theorem that follows, a connection is made between the Stein operator and a derivative of the KL divergence with respect to the perturbation magnitude ϵ of the transform $x \mapsto \boldsymbol{T}(x) = x + \epsilon\boldsymbol{\phi}(x)$. In the theorem we make the dependence on the function $\boldsymbol{\phi}$ explicit for $\boldsymbol{T}$ by writing $\boldsymbol{T}_{\boldsymbol{\phi}}$.

Theorem 3.8. *Consider $x \mapsto \boldsymbol{T}_{\boldsymbol{\phi}}(x) = x + \epsilon\boldsymbol{\phi}(x)$ and let $q_{[\boldsymbol{T}]}$ be the density of $Z = \boldsymbol{T}(X)$ when X has density q. Then we have:*

$$\nabla_\epsilon KL(q_{[\boldsymbol{T}_{\boldsymbol{\phi}}]} \,||\, p)\Big|_{\epsilon=0} = -\,\mathrm{E}_{X\sim q}[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(X))],$$

where $\boldsymbol{\phi} \mapsto \mathcal{A}_p\boldsymbol{\phi} = (\nabla \log p)\boldsymbol{\phi}^T + \nabla\boldsymbol{\phi}$ is the Stein operator.

Proof. See Appendix A.2. □

The remarkable observation to make now is that $\boldsymbol{\phi}^*_{q,p}$ from equation (17) is such that the following holds:

$$\frac{\boldsymbol{\phi}^*_{q,p}}{||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d}} = \arg\max\left\{-\nabla_\epsilon KL(q_{[\boldsymbol{T}_{\boldsymbol{\phi}}]} \,||\, p)\Big|_{\epsilon=0} \;\middle|\; \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\right\}.$$

In other words, the function $\boldsymbol{\phi}^*_{q,p}$, being the optimal solution (direction) for the KSD, turns out to be equal to the direction that yields the steepest descent of KL divergence of all functions $\boldsymbol{\phi} \in \mathcal{H}^d$ such that $||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1$.

Lemma 3.9 (Lemma from Liu and D. Wang 2016)**.** *Assume that the same conditions as in Theorem 3.7 and Theorem 3.8 hold. Consider every possible perturbation (direction) $\boldsymbol{\phi}$ in the ball $\mathcal{B} = \{\boldsymbol{\phi} \in \mathcal{H}^d \mid \quad ||\boldsymbol{\phi}||^2_{\mathcal{H}^d} \leq S(q,p)\}$ in RKHS $\mathcal{H}^d$. The direction of steepest descent in $\mathcal{B}$, i.e. the direction that maximizes $-\nabla_\epsilon KL(q_{[\boldsymbol{T}]} \,||\, p)\Big|_{\epsilon=0}$ is $\boldsymbol{\phi}^*_{q,p}$ from equation (17). This choice of perturbation direction results in the following equality:*

$$\nabla_\epsilon KL(q_{[\boldsymbol{T}^*]} \,||\, p)\Big|_{\epsilon=0} = -S(q,p),$$

where $\boldsymbol{T}^*$ is the mapping $x \mapsto \boldsymbol{T}^*(x) = x + \epsilon\boldsymbol{\phi}^*_{q,p}(x)$.

Proof. By definition, we have $S(q,p) = \left\{[\mathrm{E}_{x\sim q}(\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(x)))]^2 \;\middle|\; \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\right\}$ and the result of Theorem 3.7 gives a specific value to it: $S(q,p) = ||\boldsymbol{\phi}^*_{q,p}||^2_{\mathcal{H}^d}$. Furthermore, Theorem 3.7 also gives a useful identity that we are going to use below: $\langle\boldsymbol{\phi}, \boldsymbol{\phi}^*_{q,p}\rangle_{\mathcal{H}^d} = \mathrm{E}_q[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi})]$ for any $\boldsymbol{\phi} \in \mathcal{H}^d$. We can apply the result of Theorem 3.8 to all functions in $\mathcal{B}$. This way, $\phi^*_{q,p} \in \mathcal{B}$ and we have:

$$\begin{aligned}\max\left\{-\nabla_\epsilon KL(q_{[\boldsymbol{T}_{\boldsymbol{\phi}}]} \,||\, p)\Big|_{\epsilon=0} \;\middle|\; \boldsymbol{\phi} \in \mathcal{B}\right\} &= \max\left\{\mathrm{E}_{X\sim q}[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(X)) \mid \boldsymbol{\phi} \in \mathcal{B}\right\} \\ &= \max\left\{\langle\boldsymbol{\phi}, \boldsymbol{\phi}^*_{q,p}\rangle_{\mathcal{H}^d} \mid \boldsymbol{\phi} \in \mathcal{B}\right\}. \end{aligned} \tag{21}$$

For $\boldsymbol{\phi} \in \mathcal{B}$ we have by the Cauchy-Schwarz inequality that:

$$|\langle \boldsymbol{\phi}, \boldsymbol{\phi}^*_{q,p} \rangle_{\mathcal{H}^d}| \leq ||\boldsymbol{\phi}||_{\mathcal{H}^d} ||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d} \leq \sqrt{S(q,p)} ||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d}.$$

Let us pick $\tilde{\boldsymbol{\phi}} = \frac{\phi^*_{q,p}\sqrt{S(q,p)}}{||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d}}$. This gives $\langle \tilde{\boldsymbol{\phi}}, \boldsymbol{\phi}^*_{q,p} \rangle_{\mathcal{H}^d} = ||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d} \sqrt{S(q,p)}$. Hence, we have upper bounded the maximum in equation (21) and also attained this upper bound. The maximum is therefore equal to the upper bound. This gives:

$$\begin{aligned} \max \left\{ -\nabla_\epsilon KL(q_{[\boldsymbol{T}_{\boldsymbol{\phi}}]} \,||\, p)\Big|_{\epsilon=0} \;\Big|\; \boldsymbol{\phi} \in \mathcal{B}] \right\} &= \sqrt{S(q,p)} ||\boldsymbol{\phi}^*_{q,p}||_{\mathcal{H}^d} \\ &= \sqrt{S(q,p)}\sqrt{S(q,p)} \\ &= S(q,p). \end{aligned}$$

□

Lemma 3.9 can be used to construct a procedure to transform a reference density q_0 to the target density p, as the lemma gives the direction of steepest descent in $\mathcal{B}$, i.e. the direction that maximizes $-\nabla_\epsilon KL(q_{[\boldsymbol{T}]} \,||\, p)\Big|_{\epsilon=0}$. This direction is given by $\boldsymbol{\phi}^*_{q,p}$. After every iteration of the transformation function $\boldsymbol{T}$, using $\boldsymbol{\phi}^*_{q,p}$ to create this transformation function $\boldsymbol{T}$, the KL divergence between q and p shrinks. The goal is to let q resemble p as closely as possible.

The procedure is as follows:

Algorithm 1 Iterative procedure for the transformation of densities

Input: A target density p, an initial density q_0 and a sequence of step-sizes $\{\epsilon_\ell\}_{\ell \geq 0}$.
Output: A sequence of densities $\{q_\ell\}_{\ell \geq 0}$ that becomes an approximation of the density p.

1: **for** $\ell = 0, 1, 2, \ldots$ **do**
2: Compute $\boldsymbol{\phi}^*_{q_\ell,p}$ using equation (17).
3: Compute $\boldsymbol{T}^*_\ell(x) = x + \epsilon_\ell \boldsymbol{\phi}^*_{q_\ell,p}(x)$.
4: Let $q_{\ell+1} = q_{\ell[\boldsymbol{T}^*_\ell]}$, according to equation (20).
5: **end for**

The idea behind this procedure is as follows: the initial density q_0 induces an initial transform $x \mapsto \boldsymbol{T}^*_0(x) = x + \epsilon_0 \boldsymbol{\phi}^*_{q_0,p}(x)$. In turn, this transformation induces a new density $q_1 = q_{0[\boldsymbol{T}_0]}$. This transformation approximately reduces KL divergence between q_1 and p by $\epsilon_0 S(q_1, p)$ for a small step-size ϵ_0. To continue, a new transformation is created as $x \mapsto \boldsymbol{T}^*_1(x) = x + \epsilon_1 \boldsymbol{\phi}^*_{q_1,p}(x)$. As before, this transforms q_1 into q_2 and (again) approximately decreases the KL divergence between q_2 and p by $\epsilon_1 S(q_1, p)$. This procedure results in a sequence of densities $\{q_\ell\}_{\ell \geq 1}$ for iterations $\ell = 1, 2, \ldots$. Ultimately, it all revolves around

$$q_{\ell+1} = q_{\ell[\boldsymbol{T}^*_\ell]}, \tag{22}$$

$$\text{with } x \mapsto \boldsymbol{T}^*_\ell(x) = x + \epsilon_\ell \boldsymbol{\phi}^*_{q_\ell,p}(x), \tag{23}$$

for small enough perturbation sizes $\{\epsilon_\ell\}_{\ell \geq 1}$. The ideal scenario of this iterative scheme is to eventually converge to the target density p. This would mean that $\boldsymbol{\phi}^*_{q_\infty,p}$ is the zero map and

$\boldsymbol{T}_\infty^*$ is equal to the identity map. It should be remembered that $q = p$ if and only if $\boldsymbol{\phi}_{q,p}^*$ is the zero map, see e.g. equation (18) and the explanation below Theorem 3.7.

Almost all ingredients are now available to introduce Stein variational gradient descent, but first it is necessary to study how the iterative scheme of equations (22) and (23) can be approximated. The reason an approximation is needed is that the expectation to calculate $\boldsymbol{\phi}_{q,p}^*$ has to be evaluated and this involves a gradient of the target distribution p. Furthermore, it is computationally hard to track density functions, as functions are infinite dimensional. Hence, every density q_i is replaced by a sample from it.

Let us first describe the practical steps and then state the algorithm. A set of M particles is drawn, denoted $\{x_i^0\}_{i=1}^M$ with the superscript denoting that these particles are drawn from the density q_0. Next, the iterative scheme of equations (22) and (23) has to be carried out. To calculate the necessary ingredients, $\boldsymbol{T}_\ell^*$ is needed in the ℓ-th iteration and this requires $\boldsymbol{\phi}_{q_\ell,p}^*(\cdot) = \mathrm{E}_{X\sim q_\ell}[\mathcal{A}_p k(X,\cdot)]$. This is an expectation with respect to q_ℓ and it is approximated by an empirical mean over the particles that are available in iteration ℓ, i.e. the set $\{x_i^\ell\}_{i=1}^M$.

We can now give the SVGD algorithm in Algorithm 2. Originally, the authors called this method 'Bayesian inference via variational gradient descent', but we adopt the naming 'Stein variational gradient descent'.

Algorithm 2 Pseudocode for Stein variational gradient descent (Liu and D. Wang 2016)

Input: A target density p, initial particles $\{x_i^0\}_{i=1}^M$ and a sequence of step-sizes $\{\epsilon_\ell\}_{\ell\geq 0}$.
Output: A set of particles $\{x_i\}_{i=1}^M$ that resembles the distribution with density p.

for $\ell = 0, 1, 2, \ldots$ **do**
  for $i = 1, \ldots, M$ **do**
    $x_i^{\ell+1} \leftarrow x_i^\ell + \epsilon_\ell \widehat{\boldsymbol{\phi}^*}(x_i^\ell)$, with $\widehat{\boldsymbol{\phi}^*}(x) = \frac{1}{M}\sum_{j=1}^M [k(x_j^\ell, x)\nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x)]$.
  end for
end for

Remark 3.10. The function $\widehat{\boldsymbol{\phi}^*}$ is not normalized, as this implementation of SVGD uses the results presented in Lemma 3.9.

This algorithm gives a way to iteratively transform the initial particles $\{x_i^0\}_{i=1}^M$ in such a way that the transformed set of particles approximates p. The algorithm acts in the same way as a gradient descent type algorithm and SVGD can be seen as a gradient descent type algorithm for the KL divergence functional. The direction of steepest descent can be interpreted as follows. First note that the algorithm lets the particles move in the direction of $\widehat{\boldsymbol{\phi}^*}(x)$. This $\widehat{\boldsymbol{\phi}^*}$ is composed of two gradient terms and that is also the reason it is a gradient descent type of algorithm. The first term in $\widehat{\boldsymbol{\phi}^*}$ consists of the gradients of the logarithm of p, weighted by the kernel function. In fact, it is a kernel weighted average over all particles. This way the kernel smooths the gradients, but the gradient still points in the direction of areas that increase p. The second term in $\widehat{\boldsymbol{\phi}^*}$ consists of the gradients of the kernel function with respect to the particles. This term is what the authors name a repulsive force. Its aim is to prevent the particles to collapse with other particles in (local) modes of p. Let us give an example from Liu and D. Wang 2016.

Example 3.11. Consider the radial basis function (RBF) kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by $(x, x') \mapsto k(x, x') = \exp(-\frac{1}{h}||x - x'||^2)$ for some $h > 0$. This means that we get the following:

$$\sum_{j=1}^{M} \nabla_{x_j^\ell} k(x_j^\ell, x) = \sum_{j=1}^{M} \frac{2}{h}(x - x_j^\ell) k(x_j^\ell, x). \tag{24}$$

The interpretation of this derivative is that points x_j^ℓ which are close to x have a higher value of $k(x_j^\ell, x)$, as for those points $||x_j^\ell - x||^2$ is relatively small. This means that these points are more repelled and hence $\sum_{j=1}^{M} \nabla_{x_j^\ell} k(x_j^\ell, x)$ acts as a repulsive force.

Another observation can be made by taking the case of a single particle into consideration. The case $M = 1$ reduces the algorithm to MAP for any kernel that satisfies the condition that $\nabla_{x'} k(x', x) = 0$. If we look at equation (24), then for the RBF kernel function this condition holds.

A big difference with Monte Carlo methods is that these methods rely on a large number M of particles to get a good approximation of p, as these methods take an average over all the points. Another difference is that in the presented method a (deterministic) repulsive force is used to get diverse points instead of using randomization, which is typical in Monte Carlo approximation methods. In the presented method a repulsive force is used to get non-equal particles which do not collapse into local modes of p, but Monte Carlo methods rely on randomization to make sure that particles do not all end up in the same local modes of p.

To conclude the theory, let us look at the computational implementation of Algorithm 2. The main difficulty in the algorithm is the update to calculate the term $\nabla \log p$ for all particles $\{x_i\}_{i=1}^M$. This is an even more demanding task if we have a lot of data $\mathcal{D} = \{D_j\}_{j=1}^n$. This is due to the fact that $p(x|\mathcal{D}) \propto p_0(x) \prod_{j=1}^n p(D_j|x)$ becomes more difficult to handle for larger n, simply because the product contains more terms. A solution for this is proposed in the form of an approximation of $\nabla \log p$. By simply using a subset of the original dataset, an approximation for $\nabla \log p$ is made, i.e. take $\Omega \subset \{1, 2, \ldots, n\}$ as a subset of the original dataset, then the following approximation can be made:

$$\nabla \log p(x) \approx \nabla \log p_0(x) + \frac{n}{|\Omega|} \sum_{j \in \Omega} \nabla_x p(D_j|x). \tag{25}$$

By using this approximation, it is only necessary to consider the data points indexed by Ω and not the full dataset. The last concern in the algorithm is the evaluation of the kernel function on $\{x_i\}_{i=1}^M$. In fact it becomes a kernel matrix $\{k(x_i, x_j)\}_{ij}$ and evaluation of this matrix is $\mathcal{O}(M^2)$. If it is necessary to use a large M (which in practice is not always necessary, see e.g. Liu and D. Wang 2016), then a similar technique as for $\nabla \log p$ can be used, i.e. subsampling $\bar{\Omega} \subset \{1, 2, \ldots, M\}$ to approximate the sum $\sum_{i=1}^M k(x_i, x)$ by $\frac{M}{|\bar{\Omega}|} \sum_{i \in \bar{\Omega}} k(x_i, x)$.

In the last part of the original SVGD paper Liu and D. Wang 2016, the authors perform a test of Algorithm 2 on a toy example to show its workings. For the numerical experiments, the RBF kernel is used with the bandwidth $h = \text{med}^2 / \log M$, where med is the median of the pairwise distances between the current points $\{x_i^\ell\}_{i=1}^M$ in iteration ℓ. Hence, in every iteration the RBF kernel k_h is changed. In what follows, we describe the Gaussian mixture distribution experiment. The target density is given as $x \mapsto p(x) = \frac{1}{3}\varphi(x; -2, 1) + \frac{2}{3}\varphi(x; 2, 1)$ and as initial density $x \mapsto q_0(x) = \varphi(x; -10, 1)$ is used, where $x \mapsto \varphi(x : \mu, \sigma^2)$ denotes the density of a $\mathcal{N}(\mu, \sigma^2)$ random variable. In Figure 6 the evolution of the distribution of particles is shown after an increasing number of iterations. Initially, the density q_0, from which our first $M = 100$

particles are sampled, does not resemble the target density. It is the case that after 500 iterations the evolved particles follow the target density quite closely.

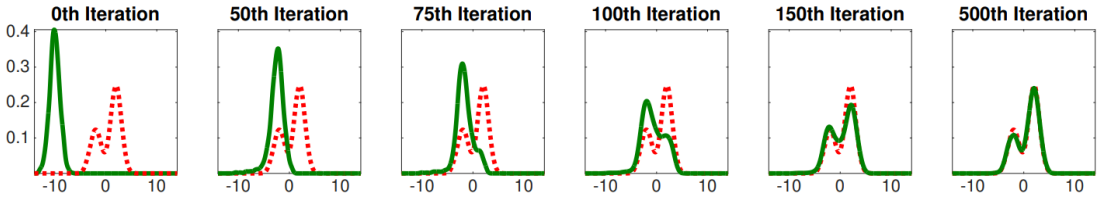

Figure 6: Toy example with Gaussian mixture distribution. The red dashed lines are the target density function and the solid green lines are the densities of the particles at different iterations of the SVGD algorithm (estimated using a kernel density estimator). Note that the initial density is set to have almost zero overlap with the target density, but the method demonstrates the ability to recover the full target density. $M = 100$ particles are used. Picture and caption from Liu and D. Wang 2016.

4 Repulsive deep ensembles are Bayesian

In this section we continue where we left off in Section 2, namely that we want to to overcome the pitfalls of deep ensembles in the sense that they have no means to prevent that the ensemble members end up with the same parameters. Hence, we want to find a way to counteract this problem. Inspired by SVGD (Liu and D. Wang 2016), a repulsive component in the training of the ensemble members is introduced. This repulsive component is integrated by means of a kernel function that models a repulsive action if two ensemble members are close to each other in parameter space. This prevents that the different ensemble members end up with the same parameters. We will follow this idea, proposed in D'Angelo and Fortuin 2021.

Let us recall the training procedure for ensembles of Bayesian neural networks. The data is given as $\mathcal{D}$ and is a set of i.i.d. observations. The Bayesian neural networks are usually trained by means of maximum a posteriori (MAP) estimation. The non-convexity of the MAP estimation (or optimization) problem is used by the deep ensembles to form M independently trained (and ideally also different) parameter solutions. Consider M parameter weights of NNs in an ensemble, $\{\theta_i\}_{i=1}^M$ with $\theta_i \in \mathbb{R}^d \; \forall i = 1, \dots, M$. The evolution of the parameters of the ensemble members under the gradient of the log-posterior gives the following update rule at iteration $\ell \in \mathbb{N}$:

$$\theta_i^{\ell+1} \leftarrow \theta_i^\ell + \epsilon_\ell \boldsymbol{\phi}(\theta_i^\ell), \quad \forall i = 1, \dots, M, \tag{26}$$

$$\text{with } \boldsymbol{\phi}(\theta_i^\ell) = \nabla_\theta \log p(\theta|\mathcal{D})|_{\theta=\theta_i^\ell}, \tag{27}$$

with (small) step size ϵ_ℓ. Assume that we have a stationary kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. A stationary kernel has the property that $k(\theta, \theta') = k(\theta + a, \theta' + a) \; \forall a \in \mathbb{R}^d$. In this way, the kernel function induces a homogeneous notion of distance in input space, meaning that it is independent of the exact place in input space. See e.g. Remes et al. 2017 for more details. A repulsive term can be parameterised through the gradient of the kernel:

$$\boldsymbol{\phi}(\theta_i^\ell) = \nabla_{\theta_i^\ell} \log(p(\theta_i^\ell|\mathcal{D})) - \mathcal{R}(\{\nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)\}_{j=1}^M), \quad \forall i = 1, \dots, M, \tag{28}$$

with $\mathcal{R}(\cdot)$ some general function that captures a repulsive action between the ensemble members $\{\theta_i\}_{i=1}^M$. We will not yet give $\mathcal{R}(\cdot)$ a precise form, but we will work towards it. At this point, we only give its argument: $\{\nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)\}_{j=1}^M$.

Example 4.1. To get a feeling for the repulsive term and the gradients of the kernel function, let us consider the well-known radial basis function (RBF) kernel:

$$(\theta_i, \theta_j) \mapsto k(\theta_i, \theta_j) = \exp\left(-\frac{1}{h}||\theta_i - \theta_j||^2\right), \tag{29}$$

with length scale h. Its gradient is given by:

$$\nabla_{\theta_i} k(\theta_i, \theta_j) = \frac{2}{h}(\theta_j - \theta_i) k(\theta_i, \theta_j). \tag{30}$$

This gradient ensures that θ_i gets 'repelled' from neighboring θ_j's. This is due to the negative exponential and the norm of the difference between the two weights. Let us illustrate this. Points θ_j which are close to θ_i have a higher value of $k(\theta_i, \theta_j)$, as for those points $||\theta_i - \theta_j||^2$ is relatively

small. Consider e.g. θ_1 and θ_2, with the property that $||\theta_1 - \theta_i||^2 < ||\theta_2 - \theta_i||^2$, i.e. θ_1 is closer to θ_i than θ_2 is to θ_i. This means that $k(\theta_i, \theta_1) > k(\theta_i, \theta_2)$ and these factors $k(\theta_i, \theta_1), k(\theta_i, \theta_2)$ magnify the vectors $\frac{2}{h}(\theta_1 - \theta_i)$ and $\frac{2}{h}(\theta_2 - \theta_i)$, respectively. This way, the kernel term in the kernel gradient magnifies the vector (direction) $(\theta_1 - \theta_i)$ more. In this way, a difference in the distance $||\theta_j - \theta_i||^2$ is preserved in the gradient of the kernel function $\nabla_{\theta_i} k(\theta_i, \theta_j)$. This property can be used to 'repel' particles that are close to each other. Observe that in the limit $h \to 0$ the kernel function k vanishes and hence the repulsive force also disappears. A similar example is given in Example 3.11.

To tackle the problem of inducing the same neural network functions from different parameter settings, equation (28) can be rephrased in function space instead of parameter space (D'Angelo and Fortuin 2021). Let $\boldsymbol{f} : \theta \mapsto f(\cdot\,; \theta)$ be a mapping that maps a parameter weight vector $\theta \in \Theta \subseteq \mathbb{R}^d$ to the corresponding neural network (regression) function. We let $\boldsymbol{f}_i := f(\cdot\,; \theta_i)$ denote the same neural network function, but with a certain indexed parameter weight θ_i. Take M 'particles' in function space $\{\boldsymbol{f}_i\}_{i=1}^M$ with $\boldsymbol{f}_i \in \mathcal{F}$. The 'repulsive' interaction between these particles (in function space) is modeled with some positive definite kernel k.

The implicit functional likelihood $p(y|x, \boldsymbol{f})$ is of interest. This functional likelihood is determined by the density $p(y|x, \theta)$ in weight space and the prior $p(\boldsymbol{f})$. This prior over $\boldsymbol{f}$ can be defined separately via e.g. a Gaussian process or it can be modeled as a push-forward measure of the weight-space measure $p(\theta)$. The reason behind this is that the randomness comes from the parameter weights θ and these parameters induce a NN function via $\boldsymbol{f}$. These two ingredients together, the likelihood $p(y|x, \boldsymbol{f})$ and the prior $p(\boldsymbol{f})$ yield the posterior $p(\boldsymbol{f}|x, y)$ in function space. We may abbreviate $p(\boldsymbol{f}|x, y)$ as $p(\boldsymbol{f}|\mathcal{D})$. The update rule at iteration $\ell \in \mathbb{N}$ in function space is:

$$\boldsymbol{f}_i^{\ell+1} \leftarrow \boldsymbol{f}_i^{\ell} + \epsilon_\ell \boldsymbol{\phi}(\boldsymbol{f}_i^{\ell}), \quad \forall i = 1, \ldots, M, \tag{31}$$

$$\text{with } \boldsymbol{\phi}(\boldsymbol{f}_i^{\ell}) = \nabla_{\boldsymbol{f}_i^{\ell}} \log p(\boldsymbol{f}_i^{\ell}|\mathcal{D}) - \mathcal{R}(\{\nabla_{\boldsymbol{f}_i^{\ell}} k(\boldsymbol{f}_i^{\ell}, \boldsymbol{f}_j^{\ell})\}_{j=1}^M). \tag{32}$$

Calculating this update is not tractable in practice, as it involves handling infinite dimensional functions and updating them. From a numerical point of view, it is preferred to work with finite dimensional objects. This problem has to be circumvented. The first step tackles the infinite dimensionality of the function space. A projection from the function space to a subspace is used, introduced in the following definition:

Definition 4.2. For any $A \subset \mathcal{X}$, we define $\pi_A : f \mapsto \pi_A(f) = \{f(a)\}_{a \in A}$ as the canonical projection onto A.

This projection can help in the sense that whenever the kernel has to be evaluated in function space, it will instead be evaluated on the projection $k(\pi_B(f), \pi_B(f'))$ for B being a subset of the input space given by a batch of datapoints from our dataset.

The second solution is to project this update in function space into parameter space and update particles in weight space accordingly. An update step in finite dimensional parameter space is tractable in practice. Furthermore, we are interested in functions, which are NNs parameterised by weights, so knowing the parameters suffices. To this end, the Jacobian of the i-th particle can be used as a projector:

$$\boldsymbol{\phi}(\theta_i^{\ell}) = \left(\frac{\partial \boldsymbol{f}_i^{\ell}}{\partial \theta_i^{\ell}}\right) (\nabla_{\boldsymbol{f}_i^{\ell}} \log(p(\boldsymbol{f}_i^{\ell}|\mathcal{D})) - \mathcal{R}(\{\nabla_{\boldsymbol{f}_i^{\ell}} k(\pi_B(\boldsymbol{f}_i^{\ell}), \pi_B(\boldsymbol{f}_j^{\ell}))\}_{j=1}^M)). \tag{33}$$

The rationale behind this projection is as follows. Define the function h as:

$$h(\boldsymbol{f}_i^\ell(\theta_i^\ell)) := \nabla_{\boldsymbol{f}_i^\ell} \log p(\boldsymbol{f}_i^\ell|\mathcal{D}) - \mathcal{R}(\{\nabla_{\boldsymbol{f}_i^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell)\}_{j=1}^M).$$

We explicitly write $\boldsymbol{f}_i^\ell(\theta_i^\ell)$ to make the dependence of $\boldsymbol{f}_i^\ell$ on θ_i^ℓ more clear. Let us assume that $\mathcal{R}(\{\nabla_{\boldsymbol{f}_i^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell)\}_{j=1}^M)$ can be written as a gradient with respect to $\boldsymbol{f}_i^\ell$, i.e. as $\nabla_{\boldsymbol{f}_i^\ell} \tilde{\mathcal{R}}(\{\nabla_{\boldsymbol{f}_i^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell)\}_{j=1}^M)$, for some function $\tilde{\mathcal{R}}$. In this way, we can write

$$h(\boldsymbol{f}_i^\ell(\theta_i^\ell)) = \nabla_{\boldsymbol{f}_i^\ell}(\log p(\boldsymbol{f}_i^\ell|\mathcal{D}) - \tilde{\mathcal{R}}(\{\nabla_{\boldsymbol{f}_i^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell)\}_{j=1}^M)) = \nabla_{\boldsymbol{f}_i^\ell} \tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell)),$$

with $\tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell)) := \log p(\boldsymbol{f}_i^\ell|\mathcal{D}) - \tilde{\mathcal{R}}(\{\nabla_{\boldsymbol{f}_i^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell)\}_{j=1}^M)$. Now, we can observe a chain rule in $\tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell))$. By the multivariable chain rule, we have:

$$\nabla_{\theta_i^\ell} \tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell)) = \left(\frac{\partial \boldsymbol{f}_i^\ell}{\partial \theta_i^\ell}\right) \nabla_{\boldsymbol{f}_i^\ell} \tilde{h}(\boldsymbol{f}_i^\ell) = \left(\frac{\partial \boldsymbol{f}_i^\ell}{\partial \theta_i^\ell}\right) h(\boldsymbol{f}_i^\ell).$$

In this way, we can see that the projection by means of the Jacobian $\left(\frac{\partial \boldsymbol{f}_i^\ell}{\partial \theta_i^\ell}\right)$ on the term $h(\boldsymbol{f}_i^\ell)$ can be seen as a gradient with respect to θ_i^ℓ of $\tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell))$. Hence, the update step in equation (33) can be seen as $\nabla_{\theta_i^\ell} \tilde{h}(\boldsymbol{f}_i^\ell(\theta_i^\ell))$.

The update in equation (33) has the same flavour as the update performed in SVGD, in the sense that it is a weighted combination of a gradient of the log-posterior and a term depending on the gradients of a kernel function. Let us recall that a SVGD update step can be written in parameter space as:

$$\boldsymbol{\phi}(\theta_i^\ell) = \frac{1}{M} \sum_{j=1}^M \left(k(\theta_i^\ell, \theta_j^\ell) \nabla_{\theta_i^\ell} \log p(\theta_i^\ell|\mathcal{D}) + \nabla_{\theta_j^\ell} k(\theta_j^\ell, \theta_i^\ell) \right). \tag{34}$$

The observation to make here is that the gradients are averaged across all particles using the kernel matrix in the first part of the equation. If the inference in this SVGD update scheme is moved to function space, then the update rule is given (Z. Wang et al. 2019) as:

$$\phi(\theta_i^\ell) = \left(\frac{\partial \boldsymbol{f}_i^\ell}{\partial \theta_i^\ell}\right) \left(\frac{1}{M} \sum_{j=1}^M \left(k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell) \nabla_{\boldsymbol{f}_j^\ell} \log p(\boldsymbol{f}_j^\ell|\mathcal{D}) + \nabla_{\boldsymbol{f}_j^\ell} k(\boldsymbol{f}_i^\ell, \boldsymbol{f}_j^\ell) \right) \right). \tag{35}$$

The averaging of gradients using a kernel can be dangerous in high-dimensional settings, where kernel methods can suffer from the curse of dimensionality. Furthermore, in equation (34) the gradients of the log-posterior are averaged by by means of a kernel similarity in weight space, but that might be non-ideal for multi-modal posteriors. Even worse, in equation (35) the gradients of the log-posterior are averaged according to kernel similarity in function space, but then projected back using only the i-th function Jacobian. The proposed method in equation (33) does not use any averaging of the log-posterior gradients and hence aims to come closer to the true particle gradients for deep ensembles.

In equation (28) the repulsive term is introduced, but not specified in detail. It was a general function of the gradient of a kernel. In what follows, the goal is to determine the specific form of this repulsive term such that the update rule that follows from is equivalent to the discretisation of the gradient flow dynamics of the KL divergence in Wasserstein space. We need some more theory for that and we will devote the next chapter to it.

5 Towards repulsive deep ensembles in Wasserstein space

Let us take a broad starting viewpoint by considering the optimization problem of some functional $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$, with $\mathcal{P}_2(\mathcal{X})$ denoting the set of Borel probability measures on $\mathcal{X}$ with finite second moments, called the Wasserstein space. We will use $\mathcal{X} = \mathbb{R}^d$, unless stated otherwise. A formal definition will be given when more precision is need. An example of such a functional $\mathcal{F}$ can for example be the KL divergence between an approximating measure μ and the target posterior π, so we take $\mathcal{F} : \mu \mapsto \mathcal{F}(\mu) = KL(\mu||\pi)$ for $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ assumed to be fixed, but unknown to us. In this section, our goal is to construct a 'flow' of probability measures $(\mu_t)_{t\geq 0}$, starting from some initial μ_0 such that μ_t converges to the minimizer of the KL divergence with respect to π. More specifically, we model this in the following way:

$$\inf_{\mu_t \in \mathcal{P}_2(\mathbb{R}^d)} KL(\mu_t||\pi). \tag{36}$$

We assume $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. Observe that the infimum can be attained by $\mu_t = \pi$, but π is unknown and we have to evolve our flow $(\mu_t)_{t\geq 0}$ such that it evolves into π and hence minimizes the KL divergence with respect to π.

Let us make an analogy with a more familiar problem, namely a gradient flow in $\mathbb{R}^d$.

Definition 5.1 (Gradient flow)**.** A gradient flow starting at x_0 for a continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is a differentiable map $x : [0, T] \to \mathbb{R}^d$ such that the following holds:

$$\begin{cases} x(0) = x_0, \\ x'(t) = -\nabla F(x(t)), \quad \text{if } t \in (0, T). \end{cases} \tag{37}$$

The dynamics of such a gradient flow problem and the evolution in time of the trajectory $t \mapsto x(t)$ are modeled by the following ODE:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\nabla F(x). \tag{38}$$

5.1 Wasserstein space

The goal of this section is to work towards the Wasserstein gradient flow (Ambrosio et al. 2008) and derive a particle update procedure with its theoretical motivation. We first need to set the scene and give the necessary theory and we will do that in this subsection. The Wasserstein gradient flow is a flow of measures and in what follows, we will introduce this space of measures and its corresponding theory. After that, we will focus on the 'flow' part. Let us first introduce the Wasserstein space on $\mathbb{R}^d$ in the following definition.

Definition 5.2 (Wasserstein space)**.** The Wasserstein space on $\mathbb{R}^d$, denoted $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} ||x||^2 \mathrm{d}\mu(x) < \infty \right\}, \tag{39}$$

where $\mathcal{P}(\mathbb{R}^d)$ is the set of measures on $\mathbb{R}^d$ with the usual Borel sigma algebra on $\mathbb{R}^d$.

This space can be equipped with the Wasserstein-2 distance W_2:

$$W_2^2(\nu,\mu) = \inf_{s\in\Gamma(\nu,\mu)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} ||x-y||^2 \mathrm{d}s(x,y) \right\}, \tag{40}$$

with $\Gamma(\nu,\mu)$ the set of all possible joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν and μ.

Let us now give a definition of a measure of a measurable mapping $T : \mathbb{R}^d \to \mathbb{R}^d$:

Definition 5.3. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ a measurable mapping. The pushforward measure $T_\# : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ is a mapping satisfies the following condition

$$T_\#\mu(B) = \mu(T^{-1}(B)), \quad \forall B \in \mathcal{B}(\mathbb{R}^d).$$

According to this definition, we have that if $X \sim \mu$, then $T(X) \sim T_\#\mu$, i.e. the pushforward measure determines the distribution of a transformation of a random variable. This definition puts us in place to state Brenier's theorem:

Theorem 5.4 (Brenier's theorem)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mu \ll \mathcal{L}^d$, where $\mathcal{L}^d$ denotes the Lebesgue measure on $\mathbb{R}^d$. Then, there exists a measurable mapping $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ satisfying the following two properties:*

- $(T_\mu^\nu)_\#\mu = \nu$,
- $W_2^2(\mu,\nu) = ||I - T_\mu^\nu||^2_{L^2(\mu)} = \int ||x - T_\mu^\nu(x)||^2 \mathrm{d}\mu(x)$.

The interpretation of this theorem is that a mapping exists, depending on ν and μ such that the pushforward of T_μ^ν w.r.t μ is ν and this mapping is such that the Wasserstein-2 distance between the two measures ν and μ is equal to the squared $L^2(\mu)$ norm of the difference between the identity map and the mapping T_μ^ν.

5.2 Towards the continuity equation

We will set the scene to be able to state the continuity equation. Let $I = (0,T)$, for some $T > 0$ be our time interval of interest. In what follows, we will introduce a 'differential structure' on $\mathcal{P}_2(\mathcal{X})$, with $\mathcal{X} = \mathbb{R}^d$ in such a way that we can define a gradient at every point on a curve of measures in $\mathcal{P}_2(\mathcal{X})$. To this end, we need smooth curves and a notion of a gradient. We will start with analysing the property of absolute continuity for curves $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$. Then, we will introduce the notion of a metric derivative $|\mu_t'|$. We will show that for absolutely continuous curves $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$, the metric derivative coincides with solutions of the continuity equation:

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,$$

which should be interpreted in a distributional sense. We will make it more precise what it means for this equation to hold in a distributional sense. Furthermore, we will show what it means to take a partial derivative of μ_t and a divergence of $v_t\mu_t$. Let us first give a very general definition of absolutely continuous curves.

Definition 5.5 (Definition 1.1.1 in Ambrosio et al. 2008: absolutely continuous curves)**.** Let $(\mathscr{S}, d)$ be a complete metric space and let $v : I \to \mathscr{S}$ be a curve in $(\mathscr{S}, d)$. We say that v belongs to $AC^p(I;\mathscr{S})$, for $p \in [1,\infty]$, if there exists $m \in L^p(I)$ such that

$$d(v(s), v(t)) \leq \int_s^t m(r)\mathrm{d}r, \quad \forall 0 < s \leq t \leq T. \tag{41}$$

In the case $p = 1$, we name the corresponding curves absolutely continuous and we will denote the corresponding space by $AC(I; \mathscr{S})$.

Remark 5.6. The Wasserstein space on $\mathbb{R}^d$ (where $\mathbb{R}^d$ is endowed with the usual Euclidean metric) with the Wasserstein metric is a complete metric space, see e.g. Proposition 7.1.5 in Ambrosio et al. 2008.

We can now give a (general) result about the existence of the metric derivative for a curve v in $AC^p(I; \mathscr{S})$.

Theorem 5.7 (Theorem 1.1.2 in Ambrosio et al. 2008: metric derivative)**.** *Let $p \in [1, \infty]$. Then, for any curve v in $AC^p(I; \mathscr{S})$, the limit*

$$|v'|(t) := \lim_{s \to t} \frac{d(v(s), v(t))}{|s - t|} \tag{42}$$

exists for $\mathcal{L}^1$-a.e. $t \in I$, with $\mathcal{L}^1$ denoting the Lebesgue measure. We call $|v'|$ the metric derivative of the the curve v. Moreover, the function $t \mapsto |v'|(t) \in L^p(I)$ is an admissible integrand for the RHS of equation (41) and it is minimal in the following sense: $|v'|(t) \leq m(t)$ for $\mathcal{L}^1$-a.e. $t \in I$, for each function m satisfying equation (41).

The notion of absolutely continuous curves and a metric derivative can be applied to the complete metric space $(\mathcal{P}_2(\mathcal{X}), W_2)$, with $\mathcal{X} = \mathbb{R}^d$. This puts us in place to make a connection between the continuity equation and absolutely continuous curves by means of the following theorem.

Theorem 5.8 (Theorem 8.3.1 in Ambrosio et al. 2008: absolutely continuous curves and the continuity equation)**.** *Let I be an open interval in $\mathbb{R}$, let $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ be an absolutely continuous curve (i.e. $\mu \in AC(I, \mathcal{P}_2(\mathcal{X}))$) and let $|\mu'| \in L^1(I)$ be its metric derivative, given by Theorem 5.7. Then, there exists a Borel measurable vector field $v : (x, t) \mapsto v_t(x) \in \mathbb{R}^d$ such that:*

$$v_t \in L^2(\mu_t; \mathcal{X}), \quad ||v_t||_{L^2(\mu_t; \mathcal{X})} \leq |\mu'|(t) \text{ for } \mathcal{L}^1\text{-a.e. } t \in I, \tag{43}$$

and the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{44}$$

holds in the sense of distributions, i.e.

$$\int_I \int_{\mathcal{X}} (\partial_t \phi(x, t) + \langle v_t(x), \nabla_x \phi(x, t) \rangle) \mathrm{d}\mu_t(x) \mathrm{d}t = 0, \quad \forall \phi \in C_c^\infty(\mathcal{X} \times I). \tag{45}$$

Moreover, for $\mathcal{L}^1$-a.e. $t \in I$, v_t belongs to the closure in $L^2(\mu_t; \mathcal{X})$ of the subspace generated by the gradients $\nabla \phi$ with $\phi \in C_c^\infty(\mathcal{X})$. Conversely, if a weakly continuous curve $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ satisfies the continuity equation (in the sense of distributions) for some Borel velocity field $v : (x, t) \mapsto v_t(x)$ with $||v_t||_{L^2(\mu_t; \mathcal{X})} \in L^1(I)$, then $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ is absolutely continuous and $|\mu'|(t) \leq ||v_t||_{L^2(\mu_t; \mathcal{X})}$ for $\mathcal{L}^1$-a.e. $t \in I$.

Remark 5.9. Following the convention in section 5.1 of Ambrosio et al. 2008, we say that a sequence $(\mu_n) \subset \mathcal{P}(\mathcal{X})$ is weakly convergent to $\mu \in \mathcal{P}(\mathcal{X})$ as $n \to \infty$ if

$$\lim_{n\to\infty} \int_{\mathcal{X}} f(x) \mathrm{d}\mu_n(x) = \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x),$$

for every function $f \in C_b^0(\mathcal{X})$, the space of continuous and bounded real functions defined on $\mathcal{X}$. A weakly continuous curve μ_t can be understood as being continuous with respect to this notion of weak convergence.

So, given an absolutely continuous curve μ_t, it is possible to find a time-dependent vector field v_t such that $||v_t||_{L^2(\mu_t;\mathcal{X})} \leq |\mu'|(t)$ for $\mathcal{L}^1$-a.e. $t \in I$ and such that the continuity equation holds. Conversely, if a specific μ_t solves the continuity equation for some specific vector field v_t that satisfies $||v_t||_{L^2(\mu_t;\mathcal{X})} \in L^1(I)$, i.e. $\int_I ||v_t||_{L^2(\mu_t;\mathcal{X})} \mathrm{d}t < \infty$, then the curve μ_t is absolutely continuous and we have that $||v_t||_{L^2(\mu_t;\mathcal{X})} \geq |\mu'|(t)$ for $\mathcal{L}^1$-a.e. $t \in I$. The take-away from this is that whenever a curve μ_t satisfies the continuity equation, we can find a vector field that has minimal L^2 norm. Furthermore, this minimum L^2 norm is given by $|\mu'|$. The question is now whether there is a means to find a unique vector field v_t for a given absolutely continuous curve μ_t. It turns out that such a selection principle exists. The reason we try to find a unique vector field for this absolutely continuous curve is to identify this unique vector field as the tangent vector of the curve μ_t. In order for a tangent vector to exist, we also need to have the notion of a tangent plane. In what follows, we will first give the definition of such a tangent plane and then we will state the proposition that shows the favourable properties of this tangent plane.

Definition 5.10 (Definition 8.4.1 in Ambrosio et al. 2008: tangent bundle)**.** Let $\mu \in \mathcal{P}_2(\mathcal{X})$. We define the tangent bundle at μ as $\mathrm{Tan}_\mu \mathcal{P}_2(\mathcal{X}) := \overline{\{\nabla\varphi \mid \varphi \in C_c^\infty(\mathcal{X})\}}^{L^2(\mu;\mathcal{X})}$, where the overline means closure.

Now that we have a definition of the tangent bundle (also called tangent plane) at a point $\mu \in \mathcal{P}_2(\mathcal{X})$, it is interesting to see how it links to the unique vector field v_t, found for an absolutely continuous curve μ_t in Theorem 5.8. Let us give a proposition that shows that this unique vector field v_t is in the tangent plane of this absolutely continuous curve. In this way, we can view this v_t as being the tangent vector to μ_t.

Proposition 5.11 (Proposition 8.4.5 in Ambrosio et al. 2008: tangent vector to absolutely continuous curves)**.** *Let $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ be an absolutely continuous curve and let $v_t \in L^2(\mu_t;\mathcal{X})$ be such that the continuity equation holds. Then, v_t satisfies equation (43) if and only if $v_t \in \mathrm{Tan}_{\mu_t} \mathcal{P}_2(\mathcal{X})$ for $\mathcal{L}^1$-a.e. $t \in I$. The vector v_t is uniquely determined $\mathcal{L}^1$-a.e. in I by equations (43) and (44).*

Observe that, by definition of the tangent plane, it is the case that our unique tangent vector v_t is in $L^2(\mu_t;\mathcal{X})$.

5.3 Towards the gradient flow formulation

We now focus on functionals on the Wasserstein space, i.e. functions that take measures as inputs. To properly define a gradient flow on the Wasserstein space, we need certain assumptions on these functionals. For instance, in the definition of a Fréchet subdifferential, the assumption is needed that the functional $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to (-\infty, \infty]$ is proper and lower semicontinuous, with $D(|\partial\mathcal{F}|) \subset \mathcal{P}_2^r(\mathcal{X})$. Here, $\mathcal{P}_2^r(\mathcal{X})$ denotes the measures in $\mathcal{P}_2(\mathcal{X})$ that are regular. In fact, we

have that $\mathcal{P}_2^r(\mathbb{R}^d) = \mathcal{P}_2(\mathbb{R}^d)$. Let us also give the definitions of what it means for a functional to be proper and semicontinuous. Furthermore, let us also give the meaning of $D(|\partial\mathcal{F}|)$.

Remark 5.12. A functional $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to (-\infty, \infty]$ is called proper if it has a proper effective domain, meaning that $D(\mathcal{F}) \neq \emptyset$, where $D(\mathcal{F})$ is defined as follows:

$$D(\mathcal{F}) := \{\mu \in \mathcal{P}_2(\mathcal{X}) \mid \mathcal{F}(\mu) < \infty\}.$$

This functional $\mathcal{F}$ is said to be lower semicontinuous in $\mathcal{P}_2(\mathcal{X})$ if for all $\mu \in \mathcal{P}_2(\mathcal{X})$:

$$\liminf_{\nu \to \mu} \mathcal{F}(\nu) \geq \mathcal{F}(\mu),$$

where $\nu \to \mu$ should be interpreted as convergence in Wasserstein distance.

The local slope of $\mathcal{F}$ at $\mu \in D(\mathcal{F})$, denoted as $|\partial\mathcal{F}|(\mu)$, is defined as

$$|\partial\mathcal{F}|(\mu) = \limsup_{\nu \to \mu} \frac{(\mathcal{F}(\nu) - \mathcal{F}(\mu))^+}{W_2(\nu, \mu)},$$

where we can once again understand $\nu \to \mu$ as converging in Wasserstein distance and $(a)^+ := \max\{0, a\}$ for $a \in \mathbb{R}$.

We define $D(|\partial\mathcal{F}|)$, in an analogous way as we did for $D(\mathcal{F})$,
namely $D(|\partial\mathcal{F}|) := \{\mu \in \mathcal{P}_2(\mathcal{X}) \mid |\partial\mathcal{F}|(\mu) < \infty\}$.

We are now in place to give the definition of a Fréchet subdifferential.

Definition 5.13 (Definition 10.1.1 in Ambrosio et al. 2008: Fréchet subdifferential)**.** Let $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to (-\infty, \infty]$ be a functional that is proper, lower semicontinuous and has. Let $\mu \in D(|\partial\mathcal{F}|)$. We say that $\boldsymbol{\xi} \in L^2(\mu; \mathcal{X})$ belongs to the Fréchet subdifferential $\partial\mathcal{F}(\mu)$ if

$$\mathcal{F}(\nu) - \mathcal{F}(\mu) \geq \int_{\mathcal{X}} \langle \boldsymbol{\xi}(x), T_\mu^\nu(x) - x \rangle \mathrm{d}\mu(x) + o(W_2(\mu, \nu)), \quad \text{as } \nu \to \mu,$$

where T_μ^ν refers to the optimal transport map in Theorem 5.4.

We are now ready to formally state what it means to speak about a gradient flow in the Wasserstein space.

Definition 5.14 (Definition 11.1.1 in Ambrosio et al. 2008: gradient flows)**.** We say that a map $\mu_t \in AC^2_{\mathrm{loc}}((0, \infty); \mathcal{P}_2(\mathcal{X}))$ is a solution of the gradient flow equation:

$$v_t \in -\partial\mathcal{F}(\mu_t), \quad t > 0, \tag{46}$$

if its velocity vector field $v_t \in \mathrm{Tan}_{\mu_t} \mathcal{P}_2(\mathcal{X})$ belongs to the subdifferential $\partial\mathcal{F}(\mu_t)$ from Definition 5.13 of $\mathcal{F}$ at μ_t for $\mathcal{L}^1$-a.e. $t > 0$, meaning that for $\mathcal{L}^1$-a.e. $t > 0$:

$$\mathcal{F}(\nu) - \mathcal{F}(\mu_t) \geq \int_{\mathcal{X}} \langle v_t(x), T_{\mu_t}^\nu(x) - x \rangle \mathrm{d}\mu_t(x) + o(W_2(\mu_t, \nu)), \quad \text{as } \nu \to \mu_t.$$

Remark 5.15. The space $AC^2_{\text{loc}}((0,\infty);\mathcal{P}_2(\mathcal{X}))$ is the space of curves that are locally defined, i.e. $\mu \in AC^2_{\text{loc}}((0,\infty);\mathcal{P}_2(\mathcal{X}))$ if $\mu \in AC^2(I;\mathcal{P}_2(\mathcal{X}))$ for every interval $I=(a,b)$ with $0 \le a < b$. We will consider the interval $I=(0,T)$, unless specifically stated otherwise.

This definition of a gradient flow is equivalent to the following characterisation: there exists a Borel vector field v_t such that $v_t \in \text{Tan}_{\mu_t}\mathcal{P}_2(\mathcal{X})$ for $\mathcal{L}^1$-a.e. $t>0$, $||v_t||_{L^2(\mu_t)} \in L^2_{\text{Loc}}(0,\infty)$, the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{47}$$

holds in the sense of distributions and the following inclusion of the vector field is needed:

$$v_t \in -\partial\mathcal{F}(\mu_t) \quad \text{for } \mathcal{L}^1\text{-a.e. } t>0. \tag{48}$$

5.3.1 Exponential decay of the KL divergence

We will work towards a result that shows that under a certain condition on the functional under consideration we can have exponential decay of the Wasserstein distance for a gradient flow. Let us first introduce some notation.

Remark 5.16. Following section 7.1 in Ambrosio et al. 2008, the set $\Gamma(\mu,\nu)$ is the set of joint couplings/distributions between μ and ν with marginals μ and ν. The set $\Gamma_o(\mu,\nu) \subset \Gamma(\mu,\nu)$ is the convex and weakly compact set of optimal transport plans where the minimum (as in the definition of the Wasserstein distance, see equation (40)) is attained, i.e.

$$\gamma \in \Gamma_o(\mu,\nu) \iff \int_{\mathcal{X}^2} ||x_1 - x_2||^2 \mathrm{d}\gamma(x_1,x_2) = W_2^2(\mu,\nu).$$

Definition 5.17 (Definition 9.1.1 in Ambrosio et al. 2008: λ-convexity along geodesics)**.** Let $\mathcal{X}=\mathbb{R}^d$ and let $\mathcal{F}:\mathcal{P}_2(\mathcal{X}) \to (-\infty,\infty]$. Given $\lambda \in \mathbb{R}$, we say that $\mathcal{F}$ is λ-geodesically convex in $\mathcal{P}_2(\mathcal{X})$ if for every couple $\mu,\nu \in \mathcal{P}_2(\mathcal{X})$ there exists an optimal joint distribution $\boldsymbol{\mu} \in \Gamma_o(\mu,\nu)$ such that

$$\mathcal{F}(\mu_t^{1\to 2}) \le (1-t)\mathcal{F}(\mu) + t\mathcal{F}(\nu) - \frac{\lambda}{2}t(1-t)W_2^2(\mu,\nu), \quad \forall t \in [0,1],$$

where $\mu_t^{1\to 2} := (\pi_t^{1\to 2})_{\#}\boldsymbol{\mu} = ((1-t)\pi^1 + t\pi^2)_{\#}\boldsymbol{\mu}$, with π^1,π^2 being the projections onto the first and second coordinate in $\mathcal{X}^2$, respectively. Specifically, $(1-t)\pi^1 + t\pi^2$ is the map $(x,y) \mapsto (1-t)x + ty$.

Theorem 5.18 (Theorem 11.1.4 in Ambrosio et al. 2008: exponential Wasserstein decay)**.** *Let $\mathcal{F}:\mathcal{P}_2(\mathcal{X}) \to (-\infty,\infty]$ be a lower semicontinuous λ-geodesically convex functional. If $\mu_t^i : (0,\infty) \to \mathcal{P}_2(\mathcal{X}), i=1,2$, are gradient flows (in the sense of Definition 5.14) satisfying $\mu_t^i \to \mu^i$ as $t \downarrow 0$ in $\mathcal{P}_2(\mathcal{X})$, then*

$$W_2(\mu_t^1,\mu_t^2) \le e^{-\lambda t}W_2(\mu^1,\mu^2), \quad \forall t>0.$$

In particular, for any $\mu_0 \in \mathcal{P}_2(\mathcal{X})$ there is at most one gradient flow μ_t satisfying the initial condition $\mu_t \to \mu_0$ as $t \downarrow 0$.

Hence, when the functional $\mathcal{F}$ is λ-geodesically convex, it also enjoys the exponential decay property of the Wasserstein distance along its gradient flow $v_t \in -\partial\mathcal{F}(\mu_t), t > 0$. It turns out that for the KL-divergence functional $KL(\cdot|\pi)$ it is possible to deduce that this functional is λ-geodesically convex by looking at the target measure π. This property is log-concavity, which we define below.

Definition 5.19 (Definition 9.4.9 in Ambrosio et al. 2008: log-concavity of a measure)**.** We say that a Borel probability measure $\pi \in \mathcal{P}_2(\mathcal{X})$ on $\mathcal{X}$ is log-concave if for every couple of open sets $A, B \subset \mathcal{X}$ we have:

$$\log(\pi((1-t)A + tB)) \geq (1-t)\log(\pi(A)) + t\log(\pi(B)).$$

Let us now state the theorem that shows that log-concavity is equivalent with λ-geodesic convexity of the KL-divergence functional.

Theorem 5.20 (Theorem 9.4.11 in Ambrosio et al. 2008)**.** *Let $\mathcal{X} = \mathbb{R}^d$ and let $\pi \in \mathcal{P}_2(\mathcal{X})$. Then, $KL(\cdot|\pi)$ is geodesically convex in $\mathcal{P}_2(\mathcal{X})$ if and only π is log-concave.*

5.4 Continuity equation

In this section, we will work towards the Wasserstein gradient flow. We will follow Korba, Aubin-Frankowski, et al. 2021 to give more insight in the interpretation of the continuity equation. In turn, most of what is presented in Korba, Aubin-Frankowski, et al. 2021 is based on Ambrosio et al. 2008. In particular, we are interested in the link between the continuity equation and its effect on particles.

We consider the setting of Theorem 5.8, i.e. we let I be an open interval in $\mathbb{R}$ and $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ is an absolutely continuous curve. Then, we have the existence of a Borel measurable vector field $v : (x,t) \mapsto v_t(x) \in \mathbb{R}^d$ (with certain properties) such that the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{49}$$

holds in the sense of distributions, i.e.

$$\int_I \int_{\mathcal{X}} (\partial_t \phi(x,t) + \langle v_t(x), \nabla_x \phi(x,t)\rangle) \mathrm{d}\mu_t(x)\mathrm{d}t = 0, \quad \forall \phi \in C_c^\infty(\mathcal{X} \times I). \tag{50}$$

The continuity equation is used to provide a framework for sampling from a target distribution π. The idea is to model a continuous process that transports particles from some initial distribution μ_0 towards particles sampled from π. In another way, this 'transport model' can be seen as finding a family of vector fields $(v_t)_{t\in I}$, with $v_t : \mathbb{R}^d \to \mathbb{R}^d$ transporting/evolving the measure μ_t via the continuity equation.

The family of vector fields $(v_t)_{t\in I}$ induces a change in the collection $(\mu_t)_{t\in I}$. This choice of $(v_t)_{t\in I}$ should make sure that μ_t evolves into π. The continuity equation ensures that no (probability) mass gets lost in this process. This continuity equation models the flow of the measures $(\mu_t)_{t\in I}$. Every μ_t governs a distribution of, let us say, M particles/samples at time t, let us denote one such particle as x_t and it is such that $x_t \sim \mu_t$. If this measure μ_t changes over time, then the distribution of the particles also changes. So, consider a collection $(x_t)_{t\in I}$ in $\mathbb{R}^d$ with an initial point $x_0 \in \mathbb{R}^d$ such that $x_0 \sim \mu_0$ and $x_t \sim \mu_t$ for all $t \in I$. The evolution over time of $(x_t)_{t\in I}$

is governed by the vector fields $(v_t)_{t\in I}$, with $v_t \in L^2(\mu_t)$ and $v_t : \mathbb{R}^d \to \mathbb{R}^d$ for all $t \in I$. These vector fields are such that $\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t(x_t)$. Once again, μ_t, being the law of x_t at time t, changes according to the continuity equation in (49). Let us give a proposition to capture this interplay between the particles and the continuity equation.

Proposition 5.21. *Let $\mathcal{X} = \mathbb{R}^d$. Given the $(v_t)_{t\in I}$, suppose that $\forall y \in \mathbb{R}^d$ there exists a smooth and measurable solution $x(\cdot\,; y)$ (measurable in y as well) satisfying:*

$$\begin{cases} x'(t;y) &= v_t(x(t;y)), \quad t \in [0,T], \\ x(0;y) &= y. \end{cases}$$

To make the dependence on the initial point y explicit, $x(\cdot\,; y)$ is written. Let μ_0 be a given measure and let μ_t be the law of $x(t;y)$ if y has law μ_0. Then, it is the case that μ_t satisfies the continuity equation for v_t, i.e.

$$\int_I \int_{\mathcal{X}} (\partial_t \phi(x,t) + \langle v_t(x), \nabla_x \phi(x,t)\rangle) \mathrm{d}\mu_t(x)\mathrm{d}t = 0, \quad \forall \phi \in C_c^\infty(\mathbb{R}^d \times I). \tag{51}$$

Proof. Consider a test function $\phi \in C_c^\infty(\mathbb{R}^d \times I)$. By the chain rule we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(x(t;y),t) = \nabla_x \phi(x(t;y),t) \cdot x'(t;y) + \partial_t \phi(x(t;y),t).$$

Substitute $x'(t;y) = v_t(x(t;y))$ and integrating over y with respect to μ_0 gives

$$\int \frac{\mathrm{d}}{\mathrm{d}t}\phi(x(t;y),t)\mathrm{d}\mu_0(y) = \int \left(\nabla_x \phi(x(t;y),t) \cdot v_t(x(t;y)) + \partial_t \phi(x(t;y),t)\right) \mathrm{d}\mu_0(y) \tag{52}$$

$$= \int \left(\nabla_x \phi(x,t) \cdot v_t(x) + \partial_t \phi(x,t)\right) \mathrm{d}\mu_t(x), \tag{53}$$

by the definition of μ_t as pushforward measure. Observe that the LHS of equation (51) is the integral over t of the equation in (53). Taking the integral over t on the LHS of equation (52) and using Fubini's theorem to justify the change in the order of integration gives that the LHS of equation (51) is also equal to

$$\int \int \left(\frac{\mathrm{d}}{\mathrm{d}t}\phi(x(t;y),t)\right) \mathrm{d}t\mathrm{d}\mu_0(y).$$

We took the test function $\phi \in C_c^\infty(\mathbb{R}^d \times I)$ and hence there exists an interval (a,b) such that $\phi(x,t) = 0$ for all $(x,t) \in \mathbb{R}^d \times (a,b)^c$ by its compact support. Similarly, $\phi(x(t;y),t) = 0$ for all $(y,t) \in \mathbb{R}^d \times (a,b)^c$. Hence, for all $y \in \mathbb{R}^d$ we have

$$\int \frac{\mathrm{d}}{\mathrm{d}t}\phi(x(t;y),t)\mathrm{d}t = \phi(x(b;y),b) - \phi(x(b;y),b) = 0.$$

□

The preceding proposition gives a formal measure-theoretic view on the continuity equation. We can also consider a specific case in which the measure μ_t has a continuously differentiable density m_t. This gives that $\mathrm{d}\mu_t(x) = m_t(x)\mathrm{d}x$. In turn, this gives a very concrete interpretation of the continuity equation, namely

$$\partial_t m_t + \nabla \cdot (v_t m_t) = 0,$$

for ordinary (partial) derivatives $\partial_t m_t$ and $\nabla \cdot \boldsymbol{g}(x) = \sum_i \frac{\partial}{\partial x_i} g_i(x)$, with $\boldsymbol{g}(x) = [g_1(x), \ldots, g_d(x)]^T$. In our formulation $\boldsymbol{g} = v_t m_t$.

5.5 Wasserstein gradient flow and its particle updates

In this section, we study the Wasserstein gradient flow, following section 10.4 in Ambrosio et al. 2008. We focus on a special form of functionals. Specifically, the functional we will study has this form:

$$\mu \to \mathcal{F}(\mu) = \begin{cases} \int_{\mathbb{R}^d} F(x, \rho(x), \nabla\rho(x))\mathrm{d}x, & \text{if } \mu = \rho\mathcal{L}^d, \rho \in C^1(\mathbb{R}^d) \\ \infty, & \text{otherwise,} \end{cases}$$

with $\mathcal{L}^d$ denoting the d-dimensional Lebesgue measure and $\mu = \rho\mathcal{L}^d$ means that μ admits a density ρ with respect to $\mathcal{L}^d$. Let us also assume that $F : \mathbb{R}^d \times [0, \infty) \times \mathbb{R}^d \to [0, \infty)$ is a C^2 function. Furthermore, let us assume that $F(x, 0, p) = 0$ for every $x, p \in \mathbb{R}^d$. We will also only consider strictly positive densities ρ. We denote the arguments of F as $(x, z, p) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$. For this specific form of the functional $\mathcal{F}$, we define the first variation of $\mathcal{F}$ by:

$$x \mapsto \frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x) := -\nabla \cdot F_p(x, \rho(x), \nabla\rho(x)) + F_z(x, \rho(x), \nabla\rho(x)).$$

See Appendix D.3 for some motivation behind this definition. Let us now present a theorem that, for this specific type of functional, gives the form of an element belonging to the subdifferential $\partial\mathcal{F}$ of $\mathcal{F}$ from Definition 5.13 and that is also in the tangent bundle. Observe that this is exactly the condition in the definition of a gradient flow in Definition 5.14. Hence, this theorem gives us a practical way to work with a 'gradient' of the functional $\mathcal{F}$.

Theorem 5.22 (Lemma 10.4.1 from Ambrosio et al. 2008)**.** *Let $\mu = \rho\mathcal{L}^d \in \mathcal{P}_2(\mathbb{R}^d)$ with $\rho \in C_c^2(\mathbb{R}^d)$ satisfy $\mathcal{F}(\mu) < \infty$ and assume $\boldsymbol{w} \in L^2(\mu; \mathbb{R}^d)$ belongs to the subdifferential $\partial\mathcal{F}$ at μ and the tangent bundle at μ, i.e. $\boldsymbol{w} \in \partial\mathcal{F}(\mu) \cap \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$\boldsymbol{w}(x) = \nabla\frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x), \quad \textit{for } \mu\textit{-a.e. } x \in \mathbb{R}^d.$$

Definition 5.23 (Wasserstein gradient)**.** Assume the same conditions as in Theorem 5.22. The Wasserstein gradient at $\mu = \rho\mathcal{L}^d \in \mathcal{P}_2(\mathbb{R}^d)$ of the functional $\mathcal{F}$, denoted as $\nabla_{W_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows:

$$\nabla_{W_2}\mathcal{F}(\mu) := \nabla\left(\frac{\delta\mathcal{F}(\rho)}{\delta\rho}\right), \tag{54}$$

Definition 5.24 (Wasserstein gradient flow)**.** The family of measures $(\mu_t)_{t\in I}$ satisfying the continuity equation

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)), \tag{55}$$

with $\nabla_{W_2}\mathcal{F}(\mu) := \nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho} \in L^2(\mu)$ denoting the Wasserstein gradient of $\mathcal{F}$ at $\mu = \rho \mathcal{L}^d \in \mathcal{P}_2(\mathbb{R}^d)$, is called a Wasserstein gradient flow of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$.

Observe that the Wasserstein gradient flow is given by the continuity equation with a specific choice of vector fields, namely $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$. In a sense this is just name-calling and simply a specific instantiation of our general framework. Hence, also for this specific choice of vector fields we have that the $(\mu_t)_{t\in I}$ are the laws corresponding to the curve $(x_t)_{t\in I}$ with $x_0 \sim \mu_0$. Once again, this curve in $\mathbb{R}^d$ changes according to the following ODE:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -(\nabla_{W_2}\mathcal{F}(\mu_t))(x_t). \tag{56}$$

Observe that we evaluated the Wasserstein gradient at the point $x_t \in \mathbb{R}^d$ and that this is not a mistake. The reason for this is that $\nabla_{W_2}\mathcal{F}(\mu) := \nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho}$ is a mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Remark 5.25. Observe that in this formulation we make the convention that we follow the negative Wasserstein gradient (to decrease $\mathcal{F}$), whereas before we used the convention that $\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t(x_t)$, i.e. we follow the direction in which v_t increases.

5.6 The gradient flow for the KL divergence

Let us now show an example that illustrates the theory we have developed before. It is Example 11.1.2 in Ambrosio et al. 2008 and shows how we can use the developed theory for the gradient flow of Definition 5.14 for a functional $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to (-\infty, \infty]$.

We aim to find a nonnegative solution $\rho : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}$ of a continuity equation of the following sort

$$\partial_t \rho - \nabla \cdot \left(\rho \nabla \left(\frac{\delta \mathcal{F}}{\delta \rho} \right) \right) = 0, \tag{57}$$

with

$$x \mapsto \frac{\delta \mathcal{F}(\rho)}{\delta \rho}(x) := -\nabla \cdot F_p(x, \rho(x), \nabla\rho(x)) + F_z(x, \rho(x), \nabla\rho(x))$$

defined to be the first variation of the functional $\mathcal{F}$, with the underlying assumption that our $\mathcal{F}$ can be written as follows:

$$\mathcal{F}(\rho) = \int_{\mathbb{R}^d} F(x, \rho(x), \nabla\rho(x))\mathrm{d}x, \tag{58}$$

with a smooth function $F = F(x, z, p) : \mathbb{R}^d \times [0, \infty) \times \mathbb{R}^d \to \mathbb{R}$. See Appendix D.3 for a motivation of this definition of the first variation. Observe that equation (57) has a very specific structure:

$$\partial\rho + \nabla \cdot (\rho v) = 0, \qquad \text{(continuity equation)} \tag{59}$$
$$\rho v = \rho \nabla \psi, \qquad \text{(gradient condition)} \tag{60}$$
$$\psi = -\frac{\delta \mathcal{F}(\rho)}{\delta \rho}. \qquad \text{(nonlinear relation).} \tag{61}$$

We want to have nonnegative solutions ρ that satisfy the following two constraints:

$$\rho(x,t) \geq 0, \quad \int_{\mathbb{R}^d} \rho(x,t)\mathrm{d}x = 1, \quad \forall t \geq 0,$$

that also have the property of giving a finite second moment,

$$\int_{\mathbb{R}^d} |x|^2 \rho(x,t)\mathrm{d}x < \infty, \quad \forall t \geq 0.$$

This way, we can view ρ as a density. Furthermore, this density can be linked to a measure μ as follows. Let us write $\rho(x,t) = \rho_t(x)$, then we can identify the measure μ_t as $\mu_t(A) = \int_A \rho_t(x)\mathrm{d}x$, for all measurable sets A. i.e. ρ_t is the density with respect to the Lebesgue measure on $\mathbb{R}^d$. Furthermore, this identification with measures μ_t also means that the functional $\mathcal{F}$ can be seen as a functional in $\mathcal{P}_2(\mathbb{R}^d)$. This way, a smooth and positive function ρ satisfying the equations (59)-(61) can be linked to a solution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. In fact, such a smooth and nonnegative function ρ is a solution of the equations (59)-(61) if and only if the associated measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is a solution of the gradient flow equation (46) for the functional $\mathcal{F}$. Observe for instance that the continuity equation in (59) corresponds to equation (47) in the gradient flow formulation. The gradient condition (60) is linked to the tangent formulation $v_t \in \mathrm{Tan}_{\mu_t} \mathcal{P}_2(\mathcal{X})$ for the gradient flow. The nonlinear relation (61) can be linked to the condition $v_t \in -\partial\mathcal{F}(\mu_t)$, as stated in equation (48).

Let us consider the KL divergence as our functional of interest $\mathcal{F}$. In particular, we consider it with respect to a fixed target measure $\pi \in \mathcal{P}_2(\mathcal{X})$, i.e. $\mathcal{F}(\mu) := KL(\mu||\pi)$. Let us assume these two measures admit densities with respect to the Lebesgue measure, denoted ρ and p, respectively for μ and π. This way, we can write $\mathcal{F}$ as

$$\mu \mapsto \mathcal{F}(\mu) = \int_{\mathbb{R}^d} \log\left(\frac{\rho(x)}{p(x)}\right)\rho(x)\mathrm{d}x,$$

where we recognize $F(x, \rho(x), \nabla\rho(x)) = F(x, \rho(x)) = \log(\frac{\rho(x)}{p(x)})\rho(x)$ as in equation (58). We used the convention that $F = F(x,z,p) : \mathbb{R}^d \times [0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ are the arguments of F. Note that our specific F does not depend on its third argument p. This gives as first variation:

$$
\begin{aligned}
\frac{\delta \mathcal{F}(\rho)}{\delta \rho}(x) &:= F_z(x, \rho(x), \nabla\rho(x)) - \nabla \cdot F_p(x, \rho(x), \nabla\rho(x)) \\
&= F_z(x, \rho(x)) \\
&= F_z(x, u), \quad \text{substitute } u = \rho(x), \\
&= \frac{\partial}{\partial u} \log\left(\frac{u}{p(x)}\right) u \\
&= \log\left(\frac{u}{p(x)}\right) + u\frac{p(x)}{u}\frac{1}{p(x)} \\
&= \log\left(\frac{\rho(x)}{p(x)}\right) + \rho(x)\frac{p(x)}{\rho(x)}\frac{1}{p(x)}, \quad \text{substitute back } u = \rho(x), \\
&= \log\left(\frac{\rho(x)}{p(x)}\right) + 1.
\end{aligned}
$$

This calculation gives us the precise form of the first variation of $\mathcal{F}$, i.e. $x \mapsto \frac{\delta \mathcal{F}(\rho)}{\delta \rho}(x) = \log\left(\frac{\rho(x)}{p(x)}\right) + 1$. This way, we can identify the tangent vector v as $v = \nabla \frac{\delta \mathcal{F}(\rho)}{\delta \rho} = \nabla \log(\frac{\rho}{p})$. This forms the gradient flow of the KL divergence.

In what follows, we will show why we put all this effort in Section 5.5 to define a very specific type of vector field, which we called the Wasserstein gradient and its corresponding Wasserstein gradient flow. To this end, we will again consider the KL divergence as our functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to [0, \infty)$. In the terminology of Section 5.5, observe that we have just derived the Wasserstein gradient of the functional $\mu \mapsto KL(\mu||\pi)$, denoted as $\nabla_{W_2} KL(\mu||\pi)$ and that it is equal to $\nabla \log\left(\frac{\rho}{p}\right)$. Let us formalize this in a proposition:

Proposition 5.26. *Let $\mathcal{F}(\mu) = KL(\mu||\pi)$ and let $\mu = \rho\mathcal{L}^d \in \mathcal{P}_2(\mathbb{R}^d)$ with $\rho \in C_c^2(\mathbb{R}^d)$ satisfy $\mathcal{F}(\mu) < \infty$. Assume $\nabla_{W_2} KL(\mu||\pi) \in L^2(\mu; \mathbb{R}^d)$ belongs to the subdifferential $\partial\mathcal{F}$ at μ and the tangent bundle at μ, i.e. $\nabla_{W_2} KL(\mu||\pi) \in \partial\mathcal{F}(\mu) \cap \mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$
\nabla_{W_2} KL(\mu||\pi)(x) = \nabla \log\left(\frac{\rho}{p}\right)(x) \quad \text{for } \mu\text{-a.e. } x \in \mathbb{R}^d. \tag{62}
$$

Proof. Result follows from Theorem 5.22, Definition 5.23 and the calculation in Section 5.6 for the first variation of the KL divergence. □

Now that we have a specific form of the Wasserstein gradient for the KL divergence functional, we can also give the update rule of the particles, i.e. we make the Wasserstein gradient flow in equation (56) for the KL divergence:

$$
\frac{\mathrm{d}x_t}{\mathrm{d}t} = -(\nabla \log\left(\frac{\rho}{p}\right))(x_t). \tag{63}
$$

This ODE now gives a way to evolve the particles in such a way that the measure μ_t follows the Wasserstein gradient flow of the KL divergence with respect to the target measure π. In Theorem 5.18 conditions are given under which exponential convergence towards π is given. In the setting of Theorem 5.18, we take $\mu_t^2 = \pi$ for all t. This constitutes a 'constant' gradient flow

in the sense of Definition 5.14. If the KL divergence functional $\mu \mapsto KL(\mu||\pi)$ is λ-geodesically convex, then the collection of measures $(\mu_t)_{t\geq 0}$ following the Wasserstein gradient flow of the KL divergence enjoys exponential convergence:

$$W_2(\mu_t, \pi) \leq e^{-\lambda t} W_2(\mu_0, \pi), \quad \forall t > 0.$$

This property of λ-geodesic convexity of the KL divergence is equivalent to the property of the log-concavity of the target measure π, by means of Theorem 5.20.

5.7 Particle updates via the Wasserstein gradient flow

We are going back to the setting of Section 4 and we consider a collection $\{\theta_i\}_{i=1}^M$ of parameters $\theta_i \in \mathbb{R}^d$, which we call particles and a small step size ϵ_ℓ at every iteration $\ell \in \mathbb{N}$. We use $\ell \in \mathbb{N}$ to denote discrete iteration steps and t to denote continuous time.

Assume that there exist densities with respect to the Lebesgue measure for the approximating measure μ_t and the target measure π, denoted as ρ_t and p, respectively. We would like to model the Wasserstein gradient flow of the KL divergence functional to drive the evolution of the parameters of the ensemble members according to this Wasserstein gradient flow of the KL divergence functional. So we want our collection of parameter particles $\{\theta_i\}_{i=1}^M$ to evolve according to the dynamics, governed by the ODE in equation (63):

$$\frac{\mathrm{d}\theta_i^t}{\mathrm{d}t} = -\nabla \log \left(\frac{\rho_t}{p}\right)(\theta_i^t), \quad \forall i = 1, \dots, M,$$

where we used the superscript t on θ_i^t to make explicit that these particles now evolve in continuous time according to the ODE above. We would like to discretize this ODE in time to be able to simulate the evolution of the parameters in discrete time. A discretization for $\ell \in \mathbb{N}$ is given as follows:

$$\theta_i^{\ell+1} = \theta_i^\ell + \epsilon_\ell(\nabla \log p(\theta_i^\ell) - \nabla \log \rho_\ell(\theta_i^\ell)), \quad \forall i = 1. \dots, M. \tag{64}$$

or, written differently:

$$\begin{aligned} \theta_i^{\ell+1} &\leftarrow \theta_i^\ell + \epsilon_\ell \boldsymbol{\phi}(\theta_i^\ell), \quad \forall i = 1, \dots, M, \\ \text{with } \boldsymbol{\phi}(\theta_i^\ell) &= \nabla \log p(\theta_i^\ell) - \nabla \log \rho_\ell(\theta_i^\ell), \end{aligned}$$

where $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$. Remember that we know the analytical form of the gradient for the target posterior, as the score function can be calculated without knowing the normalization constant Z for the posterior density p. In general, there is no access to the analytical form of the gradient, $\nabla \log \rho_\ell$. The reason for this is that we only have a sample of M particles available and the theoretical fact that the sequence of measures $(\mu_\ell)_{\ell\in\mathbb{N}}$ evolves according to the Wasserstein gradient flow of the KL divergence. Hence, we need to approximate the approximating measures $(\mu_\ell)_{\ell\in\mathbb{N}}$, their densities $(\rho_\ell)_{\ell\in\mathbb{N}}$, or $\nabla \log \rho_\ell$ directly. In D'Angelo and Fortuin 2021, the similarity between equation (64) and equation (28) is pointed out. Equation (28) in this setting is given by:

$$\boldsymbol{\phi}(\theta_i^\ell) = \nabla \log(p(\theta_i^\ell)) - \underbrace{\mathcal{R}(\{\nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)\}_{j=1}^M)}_{\text{repulsive term}}, \quad \forall i = 1, \ldots, M, \tag{65}$$

with $\mathcal{R}(\cdot)$ some general function that captures a repulsive action. This equation also governs an update equation for the particles $\{\theta_i^\ell\}$ for all $i = 1, \ldots, M$ and $\ell \in \mathbb{N}$ with the inclusion of a repulsive term. If the repulsive term is an approximation for the gradient of the logarithm of ρ_ℓ, then the two equations are very similar. In other words, if the following approximation holds true for all $\ell \in \mathbb{N}$:

$$\mathcal{R}(\{\nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)\}_{j=1}^M) \approx \nabla \log \rho_\ell(\theta_i^\ell), \quad \forall i = 1, \ldots, M,$$

then the update in equation (28) resembles the discretisation of a Wasserstein gradient flow for the KL divergence and in this way we can give a strong theoretical motivation for the addition of a repulsive term $\mathcal{R}(\{\nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)\}$ for the particle update scheme in equation (28). Knowing that the update rule in equation (64) aims to minimize the KL divergence between the approximating measures and the target posterior (as it follows the Wasserstein gradient flow for the KL divergence functional), it also shows that equation (28) can be seen as performing a similar task.

As argued before, there is no access to the analytical form of the gradient $\nabla \log \rho_\ell$ and hence an approximation is needed. A simple approximation in terms of a kernel function for the quantity of interest $\nabla \log \rho_\ell(\theta_i^\ell)$, based on the particles $\{\theta_i^\ell\}_{i=1}^M$ is for instance given by kernel density estimation (KDE). This KDE method gives an approximation $\hat{\rho}$ for the density ρ as $x \mapsto \hat{\rho}(x) = \frac{1}{M} \sum_{i=1}^M k(x, x_i^\ell)$, with $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a suitable kernel function. See for instance Wasserman 2006 for more details on KDE. Using this specific form of KDE, the gradient of the logarithm of this estimator $\hat{\rho}$ is given as:

$$\nabla \log \hat{\rho}(\cdot) = \frac{\sum_{i=1}^M \nabla k(\cdot, x_i)}{\sum_{i=1}^M k(\cdot, x_i)} \approx \nabla \log \rho(\cdot).$$

The latter approximation is because $\hat{\rho} \approx \rho$ and hence we also have $\nabla \log \hat{\rho} \approx \nabla \log \rho$. Using this KDE approximation in equation (64) yields the following update scheme at $\ell \in \mathbb{N}$:

$$\theta_i^{\ell+1} \leftarrow \theta_i^\ell + \epsilon_\ell \boldsymbol{\phi}(\theta_i^\ell), \quad \forall i = 1, \ldots, M, \tag{66}$$

$$\text{with } \boldsymbol{\phi}(\theta_i^\ell) = \nabla \log p(\theta_i^\ell) - \frac{\sum_{j=1}^M \nabla_{\theta_i^\ell} k(\theta_i^\ell, \theta_j^\ell)}{\sum_{j=1}^M k(\theta_i^\ell, \theta_j^\ell)}. \tag{67}$$

Comparing this formulation with equation (65), we can observe that the form of the repulsive form in equation (67) is

$$\theta \mapsto \mathcal{R}(\theta) = \frac{\sum_{j=1}^M \nabla_\theta k(\theta, \theta_j)}{\sum_{j=1}^M k(\theta, \theta_j)} = \mathcal{R}(\{\nabla_\theta k(\theta, \theta_j)\}_{j=1}^M). \tag{68}$$

Looking back at equation (28) and identifying the posterior term $p(\theta_i^\ell | \mathcal{D})$ with $p(\theta_i^\ell)$, we see that we have now identified the repulsive term in that formulation for deep ensembles. This way a

repulsive term is characterised that aims to model a repulsive force between ensemble members. See e.g. Example 4.1 for the intuition behind this repulsive force. Asymptotically, when $M \to \infty$, KDE is able to converge to the true density.

5.8 SVGD as Wasserstein gradient flow

In this section we want to put SVGD in the same perspective as we did with the Wasserstein flow in equation (56), meaning that we aim to write the (continuous) SVGD particle updates as a gradient flow problem. We mainly follow Chewi et al. 2020 and Korba, Salim, et al. 2020. Let us state equation (56) again:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -(\nabla_{W_2}\mathcal{F}(\mu_t))(x_t). \tag{69}$$

This ODE is the particle version of the Wasserstein gradient flow of the functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$. The part $\nabla_{W_2}\mathcal{F}(\mu_t)$ models the evolution of $(\mu_t)_{t\in(0,T)}$, being the laws corresponding to the curve $(x_t)_{t\in(0,T)}$ with $x_0 \sim \mu_0$. This is an exact Wasserstein gradient flow, as we really deal with the Wasserstein gradient applied to the functional $\mathcal{F}$, evaluated at μ_t. In equation (69) this Wasserstein gradient is now used to update the particles $(x_t)_{t\in(0,T)}$. The SVGD Wasserstein gradient flow is obtained by replacing the exact Wasserstein gradient by the image of this Wasserstein gradient under a kernel integral operator. Let us make this more precise.

The setting is as follows. We closely follow Korba, Salim, et al. 2020. We use the convention that $\mathcal{X} = \mathbb{R}^d$. For any $\mu \in \mathcal{P}_2(\mathcal{X})$, we define $L^2(\mu) := \{\boldsymbol{f} : \mathcal{X} \to \mathcal{X} \mid \int ||\boldsymbol{f}(x)||^2 \mathrm{d}\mu(x) < \infty\}$ and by $\langle\cdot,\cdot\rangle_{L^2(\mu)}, ||\cdot||_{L^2(\mu)}$ its inner product and norm, respectively. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a general positive definite kernel function. A function $(x, x') \mapsto k(x, x')$ is positive definite if $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0$ for any $x_1, \ldots, x_n \in \mathcal{X}, n \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathbb{R}$. A reproducing kernel Hilbert space (RKHS), denoted $\mathcal{H}$, with respect to the kernel k is the completion (with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ defined below) of the linear span of kernel functions, i.e. the completion of the set $\{f : x \mapsto f(x) = \sum_{i=1}^n a_i k(x, x_i), a_i \in \mathbb{R}, n \in \mathbb{N}, x_i \in \mathcal{X}\}$ and $\mathcal{H}$ is equipped with the inner product $\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^n a_i b_j k(x_i, x_j)$ for $x \mapsto g(x) = \sum_{j=1}^n b_j k(x, x_j)$. The space $\mathcal{H}^d$ is used to denote the space of vector functions $\boldsymbol{f} = [f_1, \ldots, f_d]^T$ with $f_i \in \mathcal{H}\ \forall i = 1, \ldots, d$. The inner product on $\mathcal{H}^d$ is $\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i\rangle_{\mathcal{H}}$, for $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}^d$. Furthermore, $||\boldsymbol{f}||^2_{\mathcal{H}^d} = \sum_{i=1}^d ||f_i||^2_{\mathcal{H}}$ for $\boldsymbol{f} \in \mathcal{H}^d$.

Take some $\mu \in \mathcal{P}_2(\mathcal{X})$. Under the assumption (see Theorem 4.26 in Steinwart and Christmann 2008) $\int k(x, x)\mathrm{d}\mu(x) < \infty$, we can specify the inclusion $\iota : \mathcal{H}^d \to L^2(\mu)$ and its adjoint $\iota^* : L^2(\mu) \to \mathcal{H}^d$, where we have the specific form of the adjoint $\iota^* := S_\mu$. Specifically, this S_μ is a kernel integral operator, induced by the kernel k and the measure μ as follows:

$$\boldsymbol{f} \mapsto S_\mu \boldsymbol{f} = \int k(x, \cdot)\boldsymbol{f}(x)\mathrm{d}\mu(x).$$

Let us also define $\mathcal{K}_\mu := \iota \circ S_\mu : L^2(\mu) \to L^2(\mu)$. Observe that $\mathcal{K}_\mu$ only differs from S_μ in its range, as $\mathcal{K}_\mu$ maps to $L^2(\mu)$ instead of $\mathcal{H}^d$. Consider functions $\boldsymbol{f} \in L^2(\mu)$ and $\boldsymbol{g} \in \mathcal{H}^d$. We have the following chain of equalities:

$$\langle \boldsymbol{f}, \iota\boldsymbol{g}\rangle_{L^2(\mu)} = \langle \iota^*\boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{H}^d} = \langle S_\mu \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{H}^d}, \tag{70}$$

where the use of the operators ι and its adjoint $\iota^* = S_\mu$ are clarified.

Remark 5.27. Given some $\mu \in \mathcal{P}_2(\mathcal{X})$, by assuming $\int k(x,x)\mathrm{d}\mu(x) < \infty$ we have that $\mathcal{H}^d \subset L^2(\mu)$. Let us remind that $||k(x,\cdot)||^2_{\mathcal{H}} = k(x,x)$. Then, for any $\boldsymbol{f} \in \mathcal{H}^d$:

$$
\begin{aligned}
||\boldsymbol{f}||^2_{L^2(\mu)} &= \int ||\boldsymbol{f}(x)||^2 \mathrm{d}\mu(x) \\
&= \int \sum_{i=1}^{d} f_i(x)^2 \mathrm{d}\mu(x), \quad \boldsymbol{f} = [f_1, \dots, f_d]^T, \\
&= \int \sum_{i=1}^{d} \langle f_i, k(x,\cdot)\rangle^2_{\mathcal{H}} \mathrm{d}\mu(x), \quad \text{reproducing property}, \\
&\leq \int \sum_{i=1}^{d} ||f_i||^2_{\mathcal{H}} ||k(x,\cdot)||^2_{\mathcal{H}} \mathrm{d}\mu(x), \quad \text{by Cauchy-Schwarz}, \\
&= \int ||\boldsymbol{f}||^2_{\mathcal{H}^d} ||k(x,\cdot)||^2_{\mathcal{H}} \mathrm{d}\mu(x) \\
&= ||\boldsymbol{f}||^2_{\mathcal{H}^d} \int k(x,x)\mathrm{d}\mu(x) \\
&< \infty.
\end{aligned}
$$

Hence, $\boldsymbol{f} \in L^2(\mu)$ and thus $\mathcal{H}^d \subset L^2(\mu)$.

In SVGD we replace the Wasserstein gradient at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, denoted $\nabla_{W_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$, by $\mathcal{K}_\mu \nabla_{W_2}\mathcal{F}(\mu)$, with the kernel integral operator $\mathcal{K}_\mu : L^2(\mu) \to L^2(\mu)$ defined earlier. The Wasserstein gradient is now not exact anymore, but 'kernelized' by a kernel integral operator $\mathcal{K}_\mu$. It leads to a general SVGD gradient flow for some functional $\mathcal{F}$ as follows:

$$
\frac{\mathrm{d}x_t}{\mathrm{d}t} = -(\mathcal{K}_{\mu_t} \nabla_{W_2}\mathcal{F}(\mu_t))(x_t). \tag{71}
$$

This kernel integral operator $\mathcal{K}_{\mu_t}$ is applied to the the Wasserstein gradient $\nabla_{W_2}\mathcal{F}(\mu_t)$ to produce a general SVGD Wasserstein gradient. Remember that $\nabla_{W_2}\mathcal{F}(\mu_t)$ is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$. Let us pick the functional $\mathcal{F}$ to be equal to the KL divergence $\mu \mapsto KL(\mu||\pi)$, because then we obtain the original SVGD Wasserstein gradient. Furthermore, this also gives a different view on how SVGD works, as for $v \in \mathcal{H}^d$ we have the following equality by using equation (70):

$$
\langle S_\mu \nabla_{W_2} KL(\mu||\pi), v\rangle_{\mathcal{H}^d} = \langle \nabla_{W_2} KL(\mu||\pi), \iota v\rangle_{L^2(\mu)}.
$$

This way, we can link $S_\mu \nabla_{W_2} KL(\mu||\pi)$ with $\nabla_{W_2} KL(\mu||\pi)$, i.e. the inner product (in $\mathcal{H}^d$) of $S_\mu \nabla_{W_2} KL(\mu||\pi)$ with a vector field $v \in \mathcal{H}^d$ is equal to the inner product (in $L^2(\mu)$) of $\nabla_{W_2} KL(\mu||\pi)$ with the inclusion operator ι applied to the same vector field $v \in \mathcal{H}^d$ (Korba, Salim, et al. 2020).

Knowing the precise form of the Wasserstein gradient for the Kullback-Leibler divergence (see Proposition 5.26) and using the operator $\mathcal{K}_{\mu_t}$ we have that

$$\mathcal{K}_{\mu_t}\nabla_{W_2}\mathcal{F}(\mu_t) = \int k(\cdot,x)\nabla\log\left(\frac{\rho_t}{p}\right)(x)\mathrm{d}\mu_t(x)$$
$$= -\int \left(k(\cdot,x)\nabla\log p(x) + \nabla_x k(x,\cdot)\right)\mathrm{d}\mu_t(x), \tag{72}$$

where in the second equality we use a partial integration result using the assumption that $\lim_{||x||\to\infty} k(\cdot,x)\rho_t(x)$ is the zero function. See Appendix C.2 for the derivation. What should be remarked now is that this kernelized gradient expression is only depending on μ_t through its (expectation) integral, i.e. we do not need the full knowledge of the distribution μ_t, but we only need to know the expectation of the integrand with respect to μ_t and that is the key to the computational feasibility of SVGD. A particle implementation of this kernelized gradient flow can be given as follows: take M initial particles $x_1^0,\dots,x_M^0$ as realisations from μ_0. Let these particles follow the (coupled system) ODE:

$$\frac{\mathrm{d}x_i^t}{\mathrm{d}t} = -\mathcal{K}_{\mu_t}\nabla\log\left(\frac{\rho_t}{p}\right)(x_i^t), \qquad \forall i = 1,\dots,M, \tag{73}$$
$$= \int \left(k(x_i^t,x)\nabla\log p(x) + \nabla_x k(x,x_i^t)\right)\mathrm{d}\mu_t(x), \qquad \forall i = 1,\dots,M. \tag{74}$$

The integral on the second line, being an expectation with respect to μ_t, can be estimated as an average over all available particles $x_1^t,\dots,x_M^t$ at time t. These particles have distribution μ_t by definition and hence the empirical average over the available particles approximates the expectation in the update rule (74). If this continuous time ODE is discretised in time, then it can be implemented numerically and the SVGD algorithm (see Algorithm 2) is obtained. For iteration $\ell\in\mathbb{N}$ we have:

$$x_i^{\ell+1} = x_i^\ell + \frac{\epsilon_\ell}{M}\sum_{j=1}^M \left(k(x_i^\ell,x_j^t)\nabla\log p(x_j^\ell) + \nabla_{x_j^\ell}k(x_j^\ell,x_i^\ell)\right), \quad \forall i = 1,\dots,M, \tag{75}$$

where ϵ_ℓ denotes some small step-size at iteration $\ell\in\mathbb{N}$. In measure space, where μ_t represents the pushforward measure of the particle x_t at time t (with $x_0\sim\mu_0$), the ODE of equation (73) (for a single particle) can be discretised in time by a gradient descent-type of approach as follows for $\ell\in\mathbb{N}$:

$$\mu_{\ell+1} = \left(\mathrm{Id} - \epsilon_\ell\mathcal{K}_{\mu_\ell}\nabla\log\left(\frac{\rho_\ell}{p}\right)\right)_{\#}\mu_\ell,$$

where $\mu_{\ell+1} = \boldsymbol{g}_{\#}\mu_l$ denotes the pushforward of the map $\boldsymbol{g}:\mathbb{R}^d\to\mathbb{R}^d$. We also let $\mathrm{Id}:\mathbb{R}^d\to\mathbb{R}^d$ be the identity map.

5.8.1 Time derivative of the KL divergence along the SVGD flow

We want to show how SVGD can dissipate the KL divergence along its gradient flow. We first give the time derivative of the KL divergence and then, by using the theory from the previous section, show the time derivative of the KL divergence for a sequence of measures $(\mu_t)_{t\geq 0}$ following the SVGD gradient flow.

Proposition 5.28. *Let $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ be a target measure with density p. Consider a collection of measures $(\mu_t)_{t\geq 0}$, with an associated collection of densities $(\rho_t)_{t\geq 0}$, satisfying a continuity equation*

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0,$$

for a collection of vector fields $(v_t)_{t\geq 0}$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t} KL(\mu_t||\pi) = \langle v_t, \nabla \log\left(\frac{\rho_t}{p}\right)\rangle_{L^2(\mu_t)}.$$

Proof. See Appendix A.4. □

Let us take a step back and also view SVGD from the point of view of the continuity equation in the following sense:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0, \quad \text{with } v_t := -\mathcal{K}_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right).$$

A well-defined and unique solution to this equation has been shown to exist, under some conditions on the kernel k and the target density p. See e.g. Lu et al. 2019. Let us assume these conditions are met. The following proposition will show that the KL divergence, along the SVGD gradient flow decreases.

Proposition 5.29 (Proposition 1 in Korba, Salim, et al. 2020)**.** *Let $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ be a target measure with density p. Consider a collection of measures $(\mu_t)_{t\geq 0}$, with a collection of corresponding densities $(\rho_t)_{t\geq 0}$, satisfying the SVGD continuity equation*

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0,$$

for a collection of vector fields $(v_t)_{t\geq 0}$ such that $v_t = -\mathcal{K}_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right)$ for all $t \geq 0$. Then we have

$$\frac{\mathrm{d}}{\mathrm{d}t} KL(\mu_t||\pi) = -||S_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right)||^2_{\mathcal{H}^d}.$$

Proof. By Proposition 5.28 we have:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} KL(\mu_t||\pi) &= \langle v_t, \nabla \log\left(\frac{\rho_t}{p}\right)\rangle_{L^2(\mu_t)} \\
&= -\langle \mathcal{K}_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right), \nabla \log\left(\frac{\rho_t}{p}\right)\rangle_{L^2(\mu_t)} \\
&= -\langle S_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right), S_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right)\rangle_{\mathcal{H}^d} \\
&= -||S_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right)||^2_{\mathcal{H}^d},
\end{aligned}$$

where in the third line we have used equation (70) with $\boldsymbol{f} = \nabla \log\left(\frac{\rho_t}{p}\right), \boldsymbol{g} = S_{\mu_t} \nabla \log\left(\frac{\rho_t}{p}\right)$ and $\mathcal{K}_{\mu_t} = \iota S_{\mu_t}$. □

6 SVGD towards convergence

In this section we work towards a convergence result for SVGD in the limit where the number of particles goes to infinity and the number of iterations goes to infinity. We will closely follow section 3 of Liu 2017. We use the same preliminaries as in Section 3.

Let consider the optimal transform $x \mapsto \boldsymbol{T}_{\mu,p}(x) = x + \epsilon \boldsymbol{\phi}^*_{\mu,p}(x)$, with $x \in \mathcal{X} = \mathbb{R}^d$ and $\boldsymbol{\phi}^*_{\mu,p}$ as in Theorem 3.7. We use p to denote the density of the target distribution ν_p, to avoid confusion with the Bayesian target posterior distribution π. Let us also define the following map $\Phi_p : \mu \mapsto (\boldsymbol{T}_{\mu,p})_{\#}\mu$, with $(\boldsymbol{T}_{\mu,p})_{\#}\mu$ denoting the pushforward measure of μ through the (measurable) transform $\boldsymbol{T}_{\mu,p}$. In other words, the function Φ_p describes the new measure of the particles after applying the transform $\boldsymbol{T}_{\mu,p}$ to the particles. This mapping fully characterizes the SVGD dynamics, as it yields the empirical measure $\hat{\mu}^M_\ell$ at iteration $\ell \in \mathbb{N}$ for M particles by recursively applying this map Φ_p, starting from our initial measure $\hat{\mu}^M_0$. More formally, Φ_p characterizes SVGD in the following way:

$$\hat{\mu}^M_{\ell+1} = \Phi_p(\hat{\mu}^M_\ell), \quad \forall \ell \in \mathbb{N} \cup \{0\}. \tag{76}$$

If the measure μ admits a density q and ϵ is small enough such that $\boldsymbol{T}_{\mu,p}$ is (locally) invertible, then the density of the measure $\mu' = \Phi_p(\mu)$, denoted q', is given by the well-known change of variables formula:

$$z \mapsto q'(z) = q(\boldsymbol{T}^{-1}_{\mu,p}(z)) \cdot |\det(\nabla \boldsymbol{T}^{-1}_{\mu,p}(z))|. \tag{77}$$

Remark 6.1. In fact, what is needed to make $\boldsymbol{T}_{\mu,p}$ (locally) invertible is that the Jacobian of $\boldsymbol{T}_{\mu,p}$ is nonsingular at a (local) point. Let us use the inverse function theorem to argue for the invertibility of $\boldsymbol{T}_{\mu,p}$. We have that $\boldsymbol{T}_{\mu,p}$ has as Jacobian $J_{\boldsymbol{T}_{\mu,p}} = \nabla \boldsymbol{T}_{\mu,p}$ at a point x: $J_{\boldsymbol{T}_{\mu,p}}(x) = I + \epsilon \nabla \boldsymbol{\phi}^*_{\mu,p}(x)$. If we can bound the spectral radius at a point x of the matrix $\nabla \boldsymbol{\phi}^*_{\mu,p}(x)$ and choose ϵ small enough, then we can make sure that the eigenvalues of $\epsilon \nabla \boldsymbol{\phi}^*_{\mu,p}(x)$ are all less than one in absolute value. Hence, the eigenvalues of $I + \epsilon \nabla \boldsymbol{\phi}^*_{\mu,p}(x)$ are in the interval $(0, 2)$, as adding the identity matrix shifts all eigenvalues by 1. In this way, if we can bound the spectral radius of $\nabla \boldsymbol{\phi}^*_{\mu,p}(x)$, the operator $\boldsymbol{T}_{\mu,p}$ can locally be be made invertible by choosing ϵ small enough (depending on the bound of the spectral radius). See Remark 6.8 for a condition such that $\nabla \boldsymbol{\phi}^*_{\mu,p}(x)$ has a bound on the spectral radius at a point x.

This map Φ_p is our tool to capture the large sample limit, i.e. the limit of $M \to \infty$ for SVGD. We will also call this the infinite particle limit. Let us assume that our initial empirical measure $\hat{\mu}^M_0$ at iteration 0 and for M particles weakly converges (as $M \to \infty$) to some limit measure μ^∞_0. This is not a strong assumption, as we are free to choose our initial measure. Let us now turn our attention to this limiting measure μ^∞_0 and apply Φ_p to it. We obtain the following recursion:

$$\mu^\infty_{\ell+1} = \Phi_p(\mu^\infty_\ell), \quad \forall \ell \in \mathbb{N} \cup \{0\}. \tag{78}$$

If we assume that $\hat{\mu}^M_0 \Rightarrow \mu^\infty_0$ at iteration 0 as $M \to \infty$, then it can be expected that $\hat{\mu}^M_\ell \Rightarrow \mu^\infty_\ell$ for all iterations $\ell \in \mathbb{N} \cup \{0\}$ if Φ_p satisfies some smoothness criterion. We used $\Rightarrow$ to denote weak convergence of a measure. A Lipschitz condition, i.e. a condition on the rate of change of a function seems appropriate, as it bounds the rate at which a function can change for any two points in its domain. In Liu 2017 the bounded Lipschitz metric is used and let us introduce it here as well.

Definition 6.2 (From Vaart and Wellner 2023)**.** Consider two measures μ and ν. Their bounded Lipschitz (BL) metric is the supremum of the difference in expectations of f with respect to μ and ν over all bounded, Lipschitz test functions $f : \mathbb{R}^d \to \mathbb{R}$:

$$BL(\mu, \nu) = \sup_f \{\mathrm{E}_\mu[f] - \mathrm{E}_\nu[f] \quad | \quad ||f||_{BL} \leq 1\}, \tag{79}$$

with $||f||_{BL} = \max\{||f||_\infty, ||f||_{\text{Lip}}\}$, where $||f||_\infty = \sup_x |f(x)|$ and $||f||_{\text{Lip}} = \sup_{x \neq y} \frac{|f(x)-f(y)|}{||x-y||_2}$. For a vector-valued bounded Lipschitz function $\boldsymbol{f} = [f_1, \ldots, f_d]^T$, its BL norm can be defined as $||\boldsymbol{f}||^2_{BL} = \sum_{i=1}^d ||f_i||^2_{BL}$.

Remark 6.3. It is the case that $BL(\mu_M, \nu) \to 0$ for $M \to \infty$ if and only if $\mu_M \Rightarrow \nu$ for $M \to \infty$. Hence, the BL metric metrizes weak convergence. See e.g. Chapter 1 in Vaart and Wellner 2023.

Let us use this strong property of the BL metric and its relation with convergence of measures by applying it to the SVGD case and the function Φ_p:

Lemma 6.4 (Lemma 3.1 in Liu 2017)**.** *Assuming* $(x, y) \mapsto \boldsymbol{g}(x, y) = (\nabla \log p(x))k(x, y) + \nabla_x k(x, y)$ *is bounded Lipschitz jointly in* (x, y) *with BL norm* $||\boldsymbol{g}||_{BL} < \infty$*, then for any two probability measures* μ *and* μ'*, we have:*

$$BL(\Phi_p(\mu), \Phi_p(\mu')) \leq (1 + 2\epsilon||\boldsymbol{g}||_{BL})BL(\mu, \mu'), \tag{80}$$

with ϵ *equal to the step-size in* $\boldsymbol{T}_{\mu,p}$*.*

Proof. See Liu 2017 for a proof. □

This is in fact the workhorse lemma for the following theorem in which we give a condition for the weak convergence of $\hat{\mu}^M_\ell$ towards μ^∞_ℓ as $M \to \infty$ for any $\ell \in \mathbb{N}$.

Theorem 6.5 (Theorem 3.2 in Liu 2017)**.** *Let* $\hat{\mu}^M_\ell$ *be the empirical measure at iteration* ℓ *for* M *particles* $\{x_i\}_{i=1}^M$ *evolving according to SVGD. Assume* $\lim_{M\to\infty} BL(\hat{\mu}^M_0, \mu^\infty_0) = 0$*. Then, for* μ^∞_ℓ *as in equation* (78)*, at any finite iteration* $\ell \in \mathbb{N}$ *it is the case that:*

$$\lim_{M\to\infty} BL(\hat{\mu}^M_\ell, \mu^\infty_\ell) = 0. \tag{81}$$

Proof. By applying the bound in Lemma 6.4 ℓ times and using the assumption about the convergence of the BL metric, i.e. $\lim_{M\to\infty} BL(\hat{\mu}^M_0, \mu^\infty_0) = 0$, gives the result. □

In Remark 6.3 we stated that if the BL metric converges to zero, then the sequence of measures also converges and that is exactly the implication of Theorem 6.5 for any finite $\ell \in \mathbb{N}$. So for any finite $\ell \in \mathbb{N}$ we have $\hat{\mu}^M_\ell \Rightarrow \mu^\infty_\ell$ as $M \to \infty$. Now that we have a proper convergence result for finite ℓ, we are interested in the limit as $\ell \to \infty$. It appears that the BL metric is not suitable for this scenario. Let us illustrate this. To be able to have convergence towards zero in the limit of $\ell \to \infty$ we need to have a factor $\alpha \in [0, 1)$ as follows: $BL(\Phi_p(\mu), \Phi(\mu')) \leq \alpha BL(\mu, \mu')$. This bound is applied ℓ times after ℓ iterations and hence, in the limit we need $\alpha^\ell \to 0$ as $\ell \to \infty$ to be able to have convergence of the BL metric. Starting from $\hat{\mu}^M_0$ with fixed M, we have that $BL(\hat{\mu}^M_\ell, \nu_p) = \mathcal{O}(\alpha^l)$. Observe that $\mathcal{O}(\alpha^\ell) \to 0$ as $\ell \to \infty$. This cannot be possible, as this would

imply that evolving SVGD for a very long time, i.e. in the limit $\ell \to \infty$, would make it possible for an empirical measure to converge to any other target measure ν_p. This can in general not be true, without making more assumptions (on for instance ν_p). Hence, there does not exist a constant $\alpha \in [0, 1)$ without any further assumptions. This means we should find a different metric to find out how SVGD evolves in the limit $\ell \to \infty$. It will turn out that the KL divergence can help us to establish convergence towards the target measure ν_p.

The useful result of Theorem 6.5 is that we do not have to start with $\hat{\mu}_0^M$, but with μ_0^∞, because this theorem shows that $\hat{\mu}_0^M$ converges to this limiting initial measure μ_0^∞ at iteration 0 as $M \to \infty$. We will assume that this limiting initial measure μ_0^∞ has a nice density and a finite KL divergence with respect to ν_p. In the theorem that follows, it is shown that the SVGD update scheme in equation (78) monotonically decreases the KL divergence between μ_ℓ^∞ and the target measure ν_p for every iteration $\ell \in \mathbb{N}$. This puts us in the position to establish the convergence $\mu_\ell^\infty \Rightarrow \nu_p$ as $\ell \to \infty$. However, we cannot do that directly via KL divergence, as we will argue after having stated the theorem.

Theorem 6.6 (Theorem 3.3 in Liu 2017)**.** *1. Assuming p is a density that satisfies Stein's identity (Lemma 3.4) $\forall \boldsymbol{\phi} \in \mathcal{H}^d$, then the measure ν_p of p is a fixed point of the map Φ_p in equation (78).*
2. Assume $R = \sup_x \{\frac{1}{2}||\nabla \log p||_{\text{Lip}} k(x,x) + 2\nabla_{xx'} k(x,x)\} < \infty$, where $\nabla_{xx'} k(x,x) = \sum_i \partial_{x_i} \partial_{x_i'} k(x,x')|_{x=x'}$, and the step size ϵ_ℓ at the ℓ-th iteration is no larger than $\epsilon_\ell^* := (2 \sup_x \rho(\nabla \boldsymbol{\phi}^*_{\mu_\ell,p} + \nabla \boldsymbol{\phi}^{*T}_{\mu_\ell,p}))^{-1}$, with $\rho(A)$ denoting the spectrum norm of the matrix A. If $KL(\mu_0^\infty||\nu_p) < \infty$ by initialisation, then

$$KL(\mu_{\ell+1}^\infty||\nu_p) - KL(\mu_\ell^\infty||\nu_p) \leq -\epsilon_\ell(1 - \epsilon_\ell R) D(\mu_\ell^\infty||\nu_p)^2, \tag{82}$$

with $D(\mu_\ell^\infty||\nu_p)$ defined to be the square root of the KSD.

Proof. See Appendix A.3. □

The interpretation of this theorem is that the population SVGD dynamics decreases the KL divergence when using sufficiently small step sizes, with a decreasing rate upper bounded by the Stein discrepancy.

Remark 6.7. The notation of the square root of kernelized Stein discrepancy (KSD), $D(\mu||\nu_p)$, used in Theorem 6.6 is different than the one used in Definition 3.6, namely $S(q,p)$. We did this to explicitly denote the dependence on the underlying measures. So we have

$$D(\mu||\nu_p) := \max\{\mathrm{E}_\mu[\mathrm{tr}(\mathcal{A}_p \boldsymbol{\phi})] \quad | \quad \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\} = \sqrt{S(q,p)}.$$

Remark 6.8. The requirement that the step size ϵ_ℓ should be smaller than ϵ_ℓ^* comes from the necessity (in the proof) that the map $\boldsymbol{T}_{\mu_\ell,p}$ has to be invertible. This can be done by invoking the inverse function theorem and bounding the spectral radius of $\nabla \boldsymbol{\phi}^*_{\mu_\ell,p}$, see Remark 6.1. Another approach is by using Lemma A.3. Let us define the Jacobian of $\boldsymbol{T}_{\mu_\ell,p}$ as $x \mapsto J_{\boldsymbol{T}_{\mu_\ell,p}}(x) = I + \epsilon \underbrace{\nabla \boldsymbol{\phi}^*_{\mu_\ell,p}(x)}_{:=B}$, then by choosing $\epsilon_\ell \leq \epsilon_\ell^* := \frac{1}{2\rho(B+B^T)}$, we get $|\det(\underbrace{I + \epsilon B}_{J_{\boldsymbol{T}_{\mu_\ell,p}}(x)})| \geq \exp(\epsilon \,\mathrm{tr}(B) - 2\epsilon^2 ||B||_F^2) > 0$, provided that $\mathrm{tr}(B)$ and $||B||_F^2$ are finite. Here, $||B||_F$ denotes the Frobenius norm of the matrix B. This result makes the Jacobian invertible for a given x, by requiring that $\epsilon_\ell \leq \epsilon_\ell^*$. Note that the inverse function theorem only locally guarantees bijectivity.

The function of interest is then guaranteed to be invertible in a local neighbourhood. However, the condition for nonsingularity of the Jacobian can be imposed globally. This can for instance be done by requiring that $\sup_x 2\rho(B + B^T)$ is bounded. Using the fact that $\rho(A) \leq ||A||$ for every square matrix A and every matrix norm $||\cdot||$ and that $||A^T||_F = ||A||_F$, we get that

$$\sup_x \rho(B + B^T) \leq \sup_x 2||B||_F = \sup_x 2||\nabla \phi^*_{\mu_\ell,p}||_F \leq \sup_x 2\sqrt{\nabla_{xx'}k(x,x)}D(\mu_\ell||\nu_p), \tag{83}$$

where the last inequality comes from equation (99) and $B = \nabla \phi^*_{\mu_\ell,p}(x)$.

The reason we cannot argue that $KL(\mu_\ell^\infty||\nu_p) \to 0$ as $\ell \to \infty$ in Theorem 6.6 is not only because $D(\mu_\ell^\infty||\nu_p)$ can be zero, but also because the right-hand side in equation (82) may to tend to zero. Hence, we only know that $KL(\mu_\ell^\infty||\nu_p)$ decreases strictly if $D(\mu_\ell^\infty||\nu_p) > 0$, but possibly not to zero. Furthermore, we arrive at a contradiction for equation (82) if we assume that $D(\mu_\ell^\infty||\nu_p)$ does not converge to zero as $\ell \to 0$. For the sake of contradiction, let us assume $D(\mu_\ell^\infty||\nu_p)$ does not converge to zero as $\ell \to \infty$. Then, we can make the right-hand side of equation (82) negative for infinitely many $\ell \in \mathbb{N}$ and hence the KL divergence between μ_ℓ^∞ and ν_p becomes negative for large enough ℓ. This is a contradiction, because the KL divergence is always nonnegative. Hence, we can conclude that $D(\mu_\ell^\infty||\nu_p) \to 0$ as $\ell \to \infty$ for a sequence of step-sizes ϵ_ℓ, but we cannot conclude that $KL(\mu_\ell^\infty||\nu_p) \to 0$ as $\ell \to \infty$.

A favourable property of the square root of the kernelized Stein discrepancy would be that $D(\mu||\nu_p) = 0$ if and only if $\mu = \nu_p$. This property can hold, but assumptions on the richness of the space $\mathcal{H}$ have to be made. In Theorem 6.6 we have that the Stein discrepancy converges to zero. In this setting, it would be ideal if we could also conclude that the measures converge (weakly). This has been studied in e.g. Gorham and Mackey 2017. Consider a sequence of measures $\{\mu_\ell\}_{\ell=1}^\infty$ and a target measure ν_p, then $D(\mu_\ell||\nu_p) \to 0$ as $\ell \to \infty$ implies that $\mu_\ell \Rightarrow \nu_p$ as $\ell \to \infty$ for the measure ν_p that is distantly dissipative and if a multi-quadric kernel is used. See Gorham and Mackey 2017 for these specific definitions. We will give more details and show more general conditions in Section 7.

If we assume that it is the case that $D(\mu_\ell^\infty||\nu_p) \to 0$ as $\ell \to \infty \implies \mu_\ell^\infty \Rightarrow \nu_p$ as $\ell \to \infty$, then Theorem 6.6 shows that for SVGD iterations it holds true that $\mu_\ell^\infty \Rightarrow \nu_p$ as $\ell \to \infty$. First, assume that the conditions in Theorem 6.5 hold and let us use it on the empirical measure $\hat{\mu}_\ell^M$. This gives that for any finite $\ell \in \mathbb{N}$ we have $\hat{\mu}_\ell^M \Rightarrow \mu_l^\infty$ as $M \to \infty$. Then, using Theorem 6.6 gives that $\hat{\mu}_\ell^M \Rightarrow \nu_p$ as first $M \to \infty$ and then $\ell \to \infty$. Hence, the two-step procedure of invoking Theorem 6.5 and then Theorem 6.6 shows the weak convergence of the empirical SVGD measure $\hat{\mu}_\ell^M$ to ν_p.

7 Convergence and tightness of SVGD measures

The aim of this section is to present results governing the convergence of SVGD. In particular, we show that it is beneficial for convergence results for SVGD if the measures, governing the SVGD particles, are forming a uniformly tight sequence of measures. In this way, a result that states that KSD metrizes weak convergence can be used and this enables us to generalise convergence results for SVGD. Furthermore, a result is shown that gives a mild condition under which the SVGD measures are in fact uniformly tight.

7.1 A motivation for tightness of measures for SVGD

Let us start by giving a motivation for studying the tightness of SVGD measures. A vital ingredient in some convergence results about SVGD, e.g. in Liu 2017 and Korba, Salim, et al. 2020 is that KSD metrizes weak convergence. As we will show in this section, this is only true in very limited cases and only holds under strict conditions. For example, it does not even hold for a very popular kernel as the Gaussian kernel. A workaround would be to show that SVGD generates a tight sequence of measures $(\mu_i)_{i\geq 1}$ that enables SVGD to satisfy the conditions under which KSD metrizes weak convergence. This result could then open the door for more general kernel functions and hence generalise convergence results for SVGD, as presented for instance in Theorem 6.6.

In what follows, we show that work in Gorham and Mackey 2017 makes explicit under which circumstances KSD is a 'good' discrepancy measure. Furthermore, a result is given that shows that KSD can fail as a discrepancy measure. We work towards a theorem that shows that SVGD, under certain assumptions, satisfies a property that enables KSD to metrize weak convergence. In fact, the convergence result of SVGD in Theorem 6.6 crucially depends on this property of KSD. However, the results in Gorham and Mackey 2017 limit the applicability of this metrizability, as their results show that it only holds for a limited choice of kernels. In this section, a result is presented that shows that it is in fact the case that KSD metrizes weak convergence for SVGD.

Let us start by recalling our definition of the (square root of) kernelized Stein discrepancy $D(\mu||\nu_p)$, where p is the density of some target measure ν_p:

$$D(\mu||\nu_p) = \max\{\mathrm{E}_\mu[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi})] \quad | \quad \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\}.$$

Remark 7.1. We will also name $D(\mu||\nu_p)$ the KSD, while strictly speaking it it is defined as the square root of the KSD as in Definition 3.6. In Appendix D.2 we give more background and unite the framework used in Gorham and Mackey 2017 with ours.

The reason we have introduced a discrepancy measure in the first place is to detect how close a certain measure is to a target measure. A favourable property of discrepancy measures is that they should also detect non-convergence towards a target measure. The following theorem shows that the KSD fails to detect when a sequence of measures for dimension $d \geq 3$ is not converging to the target, i.e. in higher dimensions we can have that $D(Q_i||\nu_p)$ converges to zero, but the sequence of 'approximating' measures $(Q_i)_{i\geq 1}$ does not converge to the target measure. We use the empirical measures as approximating measures, i.e. $Q_i = \frac{1}{i}\sum_{j=1}^{i}\delta_{x_j}$, for δ_x the Dirac measure and $x_1, \ldots, x_i \in \mathbb{R}^d$.

Theorem 7.2 (KSD fails with light kernel tails, from Gorham and Mackey 2017)**.** *Suppose $k \in C_b^{(1,1)}$ (one time continuously differentiable and uniformly bounded derivatives for both arguments). Define the kernel decay rate:*

$$\gamma(r) := \sup\{\max(|k(x,y)|, ||\nabla_x k(x,y)||_2, |\langle \nabla_x, \nabla_y k(x,y)\rangle|) \mid ||x-y||_2 \geq r\}. \tag{84}$$

If $d \geq 3$, $\nu_p = \mathcal{N}(0, I_d)$ and $\gamma(r) = o(r^{-\alpha})$ for $\alpha := (\frac{1}{2} - \frac{1}{d})^{-1}$, then there exist a sequence of measures $(Q_i)_{i\geq 1}$ such that $D(Q_i||\nu_p) \to 0$ as $i \to \infty$, while it does not imply $Q_i \Rightarrow \nu_p$ as $i \to \infty$.

Proof. See Gorham and Mackey 2017 for a proof. □

This result is far from ideal, as it limits the choice of kernels and target distributions. Furthermore, the assumptions in the theorem are satisfied by popular kernels as the RBF and Matérn kernel, see e.g. Gorham and Mackey 2017. It shows that KSD fails to detect non-convergence, even when the target measure is simply a multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. The failure of the KSD is due to its inability to enforce uniform tightness of the sequence of measures $(Q_i)_{i\geq 1}$ (Gorham and Mackey 2017).

Definition 7.3. A sequence of (arbitrary) probability measures $(\mu_i)_{i\geq 1}$ on $\mathbb{R}^d$ is uniformly tight if for every $\epsilon > 0$ there exists a number $R(\epsilon) < \infty$ such that $\limsup_{i\to\infty} \mu_i(||X||_2 > R(\epsilon)) \leq \epsilon$.

The intuition behind this definition is that no mass in the sequence of measures can escape to infinity. Furthermore, if the kernel k has (fast) decaying tails, while the score function $\boldsymbol{s}_p$ still grows, then the KSD cannot capture probability mass in the tails of the target distribution and hence it can be made arbitrarily small by a sequence of non-tight probability measures which are distributing more and more probability mass in the tails. This is the intuition behind what is happening in Theorem 7.2. Before we state a theorem which shows that, under mild conditions, KSD can detect non-converge of a sequence of arbitrary approximating probability measures with distantly dissipative target measure ν_p, let us state the definition of a distantly dissipative probability measure.

Definition 7.4 (Distant dissipativity, from Gorham and Mackey 2017)**.** A distribution ν_p (with continuously differentiable density p with support $\mathbb{R}^d$) with Lipschitz score function $\boldsymbol{s}_p := \nabla \log p$ is distantly dissipative if $\kappa_0 = \liminf_{r\to\infty} \kappa(r) > 0$ for

$$\kappa(r) = \inf\left\{ -2\frac{\langle \boldsymbol{s}_p(x) - \boldsymbol{s}_p(y), x-y\rangle}{||x-y||_2^2} \;\middle|\; ||x-y||_2 = r \right\}. \tag{85}$$

Examples of distributions being distantly dissipative are finite Gaussian mixtures with common covariance, see e.g. Gorham and Mackey 2017. We are now able to state the theorem that guarantees, under some conditions, that KSD can detect convergence of a sequence of probability measures:

Theorem 7.5 (KSD detects tight non-convergence, from Gorham and Mackey 2017)**.** *Suppose ν_p is a distantly dissipative probability measure and $(x,y) \mapsto k(x,y) = h(x-y)$, for $h \in C^2$ (twice continuously differentiable) and absolutely integrable. Assume h also has a non-vanishing generalized Fourier transform $\hat{h}$. If $(\mu_i)_{i\geq 1}$ is uniformly tight, then $D(\mu_i||\nu_p) \to 0$ as $i \to \infty$ only if $\mu_i \Rightarrow \nu_p$ as $i \to \infty$.*

Proof. See Gorham and Mackey 2017 for a proof. □

Remark 7.6. For absolutely integrable functions $f : \mathbb{R}^d \to \mathbb{R}$, i.e. $\int |f(x)|\mathrm{d}x < \infty$, the generalized Fourier transform of f is defined as $\omega \to \hat{f}(\omega) = (2\pi)^{-d/2} \int f(x) \exp(-i\langle x, \omega\rangle)\mathrm{d}x$.

It turns out that for a very specific kernel, namely the inverse multiquadric (IMQ) kernel $(x, y) \mapsto k(x, y) = (c^2 + ||x - y||_2^2)^\beta$ for some $\beta < 0$ and $c > 0$, it is the case that KSD can detect non-convergence in a broader setting (without the necessity of having tight measures). See Theorem 8 in Gorham and Mackey 2017 for more details. We are interested in a broader setting, as we do not want to restrict ourselves to solely using the IMQ kernel.

As stated before, a discrepancy measure should also detect convergence of an approximating sequence of measures $(\mu_i)_{i\geq 1}$ towards its target and this property is satisfied by KSD:

Theorem 7.7 (KSD detects convergence, from Gorham and Mackey 2017)**.** *If $k \in C_b^{(2,2)}$ (twice continuously differentiable and uniformly bounded derivatives for both arguments) and $\boldsymbol{s}_p = \nabla \log p$ is Lipschitz with $\mathrm{E}_{\nu_p}[||\boldsymbol{s}_p(X)||_2^2] < \infty$, then $D(\mu_i||\nu_p) \to 0$ as $i \to \infty$ whenever $\mu_i \Rightarrow \nu_p$ as $i \to \infty$.*

Proof. See Gorham and Mackey 2017 for a proof. □

As argued before, a vital ingredient in some convergence results about SVGD is that it is assumed that KSD metrizes weak convergence. The preceding results show that this is only true in very limited cases, i.e. the equivalence

$$D(\mu_i||\nu_p) \to 0 \text{ as } i \to \infty \iff \mu_i \Rightarrow \nu_p \text{ as } i \to \infty, \tag{86}$$

only holds under strict conditions and it does not even hold for a very popular kernel as the Gaussian kernel. To make convergence results for SVGD hold more generally, we need a workaround. A possible workaround would be to show that SVGD generates a tight sequence of measures $(\mu_i)_{i\geq 1}$ and then by means of Theorem 7.5 the only if part of this duality is satisfied. This result could then open the door for more general kernel functions and hence generalise convergence results for SVGD, as presented earlier in Theorem 6.6.

7.2 Towards a result for the tightness of measures for SVGD

This section explores a potential proof for the tightness of the measures for SVGD. We use the same notation as in Theorem 6.6 and its proof in Appendix A.3. Furthermore, we also consider the same setting and assumptions. We slightly alter the notation of the operator $\boldsymbol{T}_{\mu,p}$, where we make the dependence on the step-size more explicit, i.e. $x \mapsto \boldsymbol{T}_{\mu,\epsilon}(x) = x + \epsilon \boldsymbol{\phi}_\mu^*(x)$. Let us consider a single particle x at iteration i and denote it as x_i. In SVGD, it is updated iteratively as $x_{i+1} = \boldsymbol{T}_{\mu_i,\epsilon_i}(x_i) = x_i + \epsilon_i \boldsymbol{\phi}_{\mu_i}^*(x_i)$. We have the following chain of inequalities:

$$\begin{aligned} ||x_i||_2 &= ||x_{i-1} + \epsilon_{i-1}\boldsymbol{\phi}_{\mu_{i-1}}^*(x_{i-1})||_2 \\ &\leq ||x_{i-1}||_2 + \epsilon_{i-1}||\boldsymbol{\phi}_{\mu_{i-1}}^*(x_{i-1})||_2, \quad \text{by the triangle inequality,} \\ &\leq ||x_{i-1}||_2 + \epsilon_{i-1}\sqrt{k(x_{i-1}, x_{i-1})}D(\mu_{i-1}||\nu_p), \end{aligned} \tag{87}$$

where in the last inequality we have used the bound $||\boldsymbol{\phi}_{\mu_{i-1}}^*(x)||_2 \leq \sqrt{k(x,x)}D(\mu_{i-1}||\nu_p)$. See Lemma 7.8 for a derivation.

Lemma 7.8. *Consider the optimal the optimal update direction in $\boldsymbol{T}_{\mu,\epsilon} = x + \epsilon\boldsymbol{\phi}_\mu^*(x)$, i.e. $\boldsymbol{\phi}_{\mu_i}^* = [\phi_1, \ldots, \phi_d]^T$, with $\phi_i \in \mathcal{H}\ \forall i = 1, \ldots, d$ and $\boldsymbol{\phi}_{\mu_i}^* \in \mathcal{H}^d$, the d-dimensional RKHS with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then, the following inequality holds true for all $x \in \mathbb{R}^d$:*

$$||\boldsymbol{\phi}^*_{\mu_{i-1}}(x)||_2 \leq \sqrt{k(x,x)}D(\mu_{i-1}||\nu_p). \tag{88}$$

Proof. This proof is based on page 11 of Liu 2017. By Theorem 3.7 we have $||\boldsymbol{\phi}^*_{\mu_i}||^2_{\mathcal{H}^d} = \sum_{i=1}^d ||\phi_i||^2_{\mathcal{H}} = D(\mu_i||\nu_p)^2$. By the reproducing property of the kernel in $\mathcal{H}$ we have:

$$\phi_i(x) = \langle \phi_i, k(x,\cdot)\rangle_{\mathcal{H}}, \text{ and } k(x,x) = \langle k(\cdot,x), k(x,\cdot)\rangle_{\mathcal{H}} = ||k(x,\cdot)||^2_{\mathcal{H}} \quad \forall i = 1,\dots,d \text{ and } \forall x \in \mathcal{X}.$$

This gives the following chain of inequalities below. Consider any $x \in \mathcal{X}$, then:

$$\begin{aligned}
||\boldsymbol{\phi}^*_{\mu_i}(x)||^2_2 &= \sum_{j=1}^d \phi_j(x)^2 \\
&= \sum_{i=1}^d (\langle \phi_j, k(x,\cdot)\rangle_{\mathcal{H}})^2 \\
&\leq \sum_{j=1}^d ||\phi_j||^2_{\mathcal{H}} ||k(x,\cdot)||^2_{\mathcal{H}}, \quad \text{by Cauchy-Schwarz inequality,} \\
&= ||k(x,\cdot)||^2_{\mathcal{H}} \sum_{j=1}^d ||\phi_j||^2_{\mathcal{H}} \\
&= k(x,x)||\boldsymbol{\phi}^*_{\mu_i}||^2_{\mathcal{H}^d} \\
&= k(x,x)D(\mu_i||\nu_p)^2.
\end{aligned}$$

□

This leads us to the following lemma, which, under strong assumptions, shows that a SVGD particle has bounded norm in the iteration limit.

Lemma 7.9. *Consider the SVGD particle operator $x \mapsto \boldsymbol{T}_{\mu,\epsilon}(x) = x + \epsilon\boldsymbol{\phi}^*_{\mu}(x)$ and assume that $\sup_{i\in\mathbb{N}}\{\sqrt{k(x_i,x_i)}D(\mu_i||\nu_p)\} < \infty$. Here, μ_i denotes the SVGD pushforward measure at iteration i and x_i is a SVGD particle at iteration i. Then, we have*

$$\lim_{i\to\infty} ||x_i||_2 < \infty. \tag{89}$$

Proof. Using the inequality in equation (87) repetitively, we can get:

$$\begin{aligned}
||x_i||_2 &\leq \sum_{0\leq j\leq i-2} \epsilon_j \sqrt{k(x_j,x_j)}D(\mu_j||\nu_p) + \epsilon_{i-1}\sqrt{k(x_{i-1},x_{i-1})}D(\mu_{i-1}||\nu_p) + ||x_0||_2 \\
&= \sum_{0\leq j\leq i-1} \epsilon_j \underbrace{\sqrt{k(x_j,x_j)}D(\mu_j||\nu_p)}_{:=g_j} + ||x_0||_2.
\end{aligned}$$

Using the assumption in the lemma that $g_j < \infty \, \forall j \in \mathbb{N}$ and choosing ϵ_j suitably small and such that $\epsilon_j \to 0$ as $j \to \infty$, gives that $\sum_{0 \leq j \leq i-1} \epsilon_j g_j < \infty$ as $i \to \infty$. Let us denote this limit as M. So we have

$$||x_i||_2 \leq \sum_{0 \leq j \leq i-1} \epsilon_j g_j + ||x_0||_2 := B_i < \infty, \quad \text{as it is a finite sum.} \tag{90}$$

More interestingly, we also have, in the limit of the number of iterations $i \to \infty$, the following:

$$\lim_{i \to \infty} ||x_i||_2 \leq \lim_{i \to \infty} \sum_{0 \leq j \leq i-1} \epsilon_j g_j + ||x_0||_2 = M + ||x_0||_2 < \infty.$$

□

This lemma is the workhorse lemma for our proposition that states that the empirical measures for SVGD form a uniformly tight sequence of measures under strong assumptions.

Proposition 7.10. *Given M initial SVGD particles $x_0^1, \ldots, x_0^M$, consider the SVGD particle operator $x \mapsto \boldsymbol{T}_{\mu,\epsilon}(x) = x + \epsilon \boldsymbol{\phi}_\mu^*(x)$ and assume that $\sup_{i \in \mathbb{N}} \{\sqrt{k(x_i, x_i)} D(\mu_i || \nu_p)\} < \infty$. Here, μ_i denotes the SVGD pushforward measure at iteration i and x_i is a SVGD particle at iteration i. Furthermore, assume that $\sup_{i \in \mathbb{N}} R_i(\epsilon) < \infty$, with $R_i(\epsilon) = \sup_{j=1,\ldots,n} \{||x_i^j|| + \delta\}$ for some $\delta > 0$. Then, the empirical measure for SVGD particles generates a sequence of measures that is uniformly tight.*

Proof. Let us denote the empirical measure for SVGD as $\hat{\mu}_i^M = \frac{1}{M} \sum_{j=1}^M \delta_{x_i^j}$, at iteration i for M particles $\{x_i^j\}_{j=1}^M$. We would like to prove that this sequence of M-particle SVGD measures $(\hat{\mu}_i^M)_{i \geq 1}$ is uniformly tight. So we need that $\forall \epsilon > 0$ there exists a finite number $R(\epsilon)$ such that $\limsup_{i \to \infty} \hat{\mu}_i^M(||X||_2 > R(\epsilon)) \leq \epsilon$. As a start, let us try and bound $\hat{\mu}_i^M(||X||_2 > R_i(\epsilon))$ for a suitably chosen $R_i(\epsilon)$. We know that $\hat{\mu}_i^M = \frac{1}{M} \sum_{j=1}^M \delta_{x_i^j}$ and that $||x_i^j||_2 \leq B_i^j < \infty$ by equation (90). So, if we set $R_i(\epsilon) = \sup_{j=1,\ldots,M} \left\{ B_i^j + \delta \right\}$ for $\delta > 0$ some small number, then $\hat{\mu}_i^M(||X||_2 > R_i(\epsilon)) = 0$. Observe that $R_i(\epsilon) < \infty \, \forall i \in \mathbb{N}$. Let us set $R(\epsilon) = \sup_{i \in \mathbb{N}} R_i(\epsilon)$. Using the assumption that $R(\epsilon) < \infty$ gives that $\limsup_{i \to \infty} \hat{\mu}_i^M(||X||_2 > R(\epsilon)) = 0 \leq \epsilon$ for all $\epsilon > 0$. This shows that we have proven that $(\hat{\mu}_i^M)_{i \geq 1}$ is uniformly tight. □

A more general proposition about tightness of SVGD measures is given below. This proposition can help us in the sense that we can generalise the preceding result to make it work not only for empirical measures with a finite number of particles, but also in the infinite particle regime $M \to \infty$ for SVGD. This is for instance the setting of Theorem 6.6. Let us first give the proposition and then motivate it.

Proposition 7.11. *Let $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots$ be arbitrary measurable maps $\boldsymbol{S}_i : \mathbb{R}^d \to \mathbb{R}^d$ such that for every $x \in \mathbb{R}^d$, $\limsup_{i \to \infty} ||\boldsymbol{S}_i(x)|| < \infty$. Then, the sequence of measures $((\boldsymbol{S}_i)_{\#}\mu)_{i \geq 1}$ is uniformly tight.*

Proof. If $X \sim \mu$, then $\boldsymbol{S}_i(X) \sim (\boldsymbol{S}_i)_{\#}\mu$ by definition. Since $\sup_i ||\boldsymbol{S}_i(x)|| < \infty \, \forall x$ and $\boldsymbol{S}_i$ is measurable, $Y := \sup_i ||\boldsymbol{S}_i(X)||$ defines a finite random variable. By the inequality $||\boldsymbol{S}_i(X)|| \leq Y \, \forall i$, the uniform tightness of $((\boldsymbol{S}_i)_{\#}\mu)$ follows from the tightness of Y (in the sense that Y is finite and hence tight). □

We can identify these measurable maps $\boldsymbol{S}_1, \boldsymbol{S}_2, \dots$ as compositions of the operator $\boldsymbol{T}_{\mu_i,\epsilon_i}$ (which we abbreviate by $\boldsymbol{T}_i$) in the following sense:

particle iteration	measure
x_0	μ
$x_1 = x_0 + \epsilon_0 \boldsymbol{\phi}(x_0) = \boldsymbol{T}_0(x_0)$	$(\boldsymbol{T}_0)_{\#}\mu$
$x_2 = x_1 + \epsilon_1 \boldsymbol{\phi}(x_1) = \boldsymbol{T}_1(x_1)$	$(\boldsymbol{T}_1 \circ \boldsymbol{T}_0)_{\#}\mu$
$\vdots$	$\vdots$
$x_i = x_{i-1} + \epsilon_i \boldsymbol{\phi}(x_{i-1}) = \boldsymbol{T}_{i-1}(x_{i-1})$	$(\boldsymbol{T}_{i-1} \circ \boldsymbol{T}_{i-2} \circ \cdots \circ \boldsymbol{T}_0)_{\#}\mu$
$x_{i+1} = x_i + \epsilon_{i+1} \boldsymbol{\phi}(x_i) = \boldsymbol{T}_i(x_i)$	$(\boldsymbol{T}_i \circ \boldsymbol{T}_{i-1} \circ \cdots \circ \boldsymbol{T}_0)_{\#}\mu$

Observe that we can now identify $\boldsymbol{S}_i = \boldsymbol{T}_i \circ \boldsymbol{T}_{i-1} \circ \cdots \circ \boldsymbol{T}_0 \, \forall i \in \mathbb{N} \cup \{0\}$. In order to use Proposition 7.11, we need $\limsup_{i\to\infty} ||\boldsymbol{S}_i(x)|| < \infty \, \forall x \in \mathbb{R}^d$. If we have $\limsup_{i\to\infty} ||\boldsymbol{T}_i(x)|| < \infty \, \forall x \in \mathbb{R}^d$, then we have that this assumption satisfied, as $\boldsymbol{S}_i$ is a composition of functions $\boldsymbol{T}_i, \boldsymbol{T}_{i-1}, \dots$.

In Lemma 7.9 we have established that $\lim_{i\to\infty} ||x_i||_2 < \infty$ for an arbitrary SVGD particle x_i at iteration i. Now observe that $\lim_{i\to\infty} ||x_i||_2 = \lim_{i\to\infty} ||\boldsymbol{T}_{i-1}(x_{i-1})||_2 = \lim_{i\to\infty} ||\boldsymbol{T}_i(x_i)||_2$. This now means that $\lim_{i\to\infty} ||\boldsymbol{T}_i(x_i)||_2 < \infty$ and thus $\limsup_{i\to\infty} ||\boldsymbol{T}_i(x_i)||_2 < \infty$ for the SVGD particle x_i at iteration i. Because we took the initial particle x_0 arbitrary in our derivation and evolved it according to $\boldsymbol{T}_0, \boldsymbol{T}_1, \dots$, we can generalize the preceding result to hold for all $x \in \mathbb{R}^d$, i.e. $\limsup_{i\to\infty} ||\boldsymbol{T}_i(x)||_2 < \infty \, \forall x \in \mathbb{R}^d$. This puts us in place to use Proposition 7.11 to conclude that $((\boldsymbol{S}_i)_{\#}\mu)_{i\geq 0} = ((\boldsymbol{T}_i \circ \cdots \circ \boldsymbol{T}_0)_{\#}\mu)_{i\geq 0}$ is uniformly tight. This implies that the measures produced by SVGD are uniformly tight in the infinite particle regime $M \to \infty$.

The most crucial assumption made in this derivation is that we assumed that $\sqrt{k(x_i, x_i)} D(\mu_i || \nu_p) < \infty \, \forall i \in \mathbb{N}$. In for instance Korba, Salim, et al. 2020 a similar assumption is made, namely that $\exists C > 0$ such that $D(\mu_i || \nu_p) < C \, \forall i \in \mathbb{N}$. In a sense this seems to be a strong assumption to make, as $D(\mu || \nu_p)$ is a discrepancy measure between μ and ν_p and hence assuming this is finite for all iterations means that we stay 'finitely' close to ν_p. On the other hand, it might also be seen as a weak assumption, as it only assumes that the KSD does not become infinitely large.

7.3 A different path towards tightness of SVGD measures

In Salim et al. 2022, a different approach is used to show that the SVGD measures are uniformly tight. In that paper it is assumed that the target distribution satisfies a Talagrand-1 ($T1$) inequality. Starting from this assumption, they work towards tightness by means of a result in Dupuis and Ellis 2011:

Lemma 7.12 (Lemma 1.4.3 in Dupuis and Ellis 2011 (adapted to our needs))**.** *Let $(\mu_i)_{i\geq 1}$ be a sequence in $\mathcal{P}(\mathbb{R}^d)$ and consider a target distribution $\nu_p \in \mathcal{P}(\mathbb{R}^d)$. Assume that for each $\alpha \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \exp\langle \alpha, x \rangle \mathrm{d}\nu_p(x) < \infty \text{ and } \sup_{i\in\mathbb{N}} KL(\mu_i || \nu_p) < \infty.$$

Then, $(\mu_i)_{i\geq 1}$ *is both tight and uniformly integrable in the sense that*

$$\lim_{C\to\infty}\sup_{i\in\mathbb{N}}\int_{\{x\in\mathbb{R}^d \mid ||x||>C\}} ||x||\mathrm{d}\mu_i(x) = 0.$$

The first condition is an assumption about the target distribution and the second is about the condition that the KL divergence does not explode along the sequence. With a SVGD descent lemma for the KL divergence, we can make sure that the latter condition is satisfied, given that $KL(\mu_0||\nu_p) < \infty$. The first condition has to be assumed and the $T1$ inequality implies this condition. To be able to show this, we first need a definition:

Definition 7.13 (Talagrand's inequality $T1$)**.** The distribution ν_p satisfies the Talagrand's inequality $T1$ if there exists a $\lambda > 0$ such that for all $\mu \in \mathcal{P}(\mathbb{R}^d)$, we have $W_1(\mu, \nu_p) \leq \frac{2KL(\mu||\nu_p)}{\lambda}$. Here, $W_1(\mu, \nu_p)$ denotes the Wasserstein-1 distance between μ and ν_p.

Remark 7.14. In equation (40) the W_2 distance is defined, which differs from the W_1 distance. The W_1 distance is defined as:

$$W_1(\mu,\nu) = \inf_{s\in\Gamma(\mu,\nu)}\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} ||x-y||\mathrm{d}s(x,y)\right\},$$

with $\Gamma(\mu,\nu)$ the set of all possible joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

The assumption on ν_p, used in Salim et al. 2022 is a Talagrand-1 inequality, i.e. $T1$ in definition 7.13. In Villani et al. 2009 (Theorem 22.10) a characterisation is given for a distribution ν_p to satisfy the $T1$ inequality: ν_p satisfies $T1$ if and only if there exist $a \in \mathcal{X}$ and $\beta > 0$ such that $\int \exp(\beta||x-a||^2)\mathrm{d}\nu_p(x) < \infty$. This characterisation paves the way to the lemma. We have the following chain of inequalities. Take any $\alpha \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, then:

$$\begin{aligned}
\langle\alpha, x\rangle &\leq |\langle\alpha, x\rangle| \\
&\leq ||\alpha|| \cdot ||x||, \quad \text{by Cauchy-Schwarz,} \\
&= \left(\frac{||\alpha||}{\sqrt{\beta}}\right)\left(||x||\sqrt{\beta}\right) \\
&\leq \frac{1}{2}\frac{1}{\beta}||\alpha||^2 + \frac{1}{2}\beta||x||^2, \quad \text{because } (a-b)^2 \geq 0 \iff ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2, \\
&\leq \frac{||\alpha||^2}{2\beta} + \beta(||x-a||^2 + ||a||^2), \quad \text{as } ||x||^2 = ||x-a+a||^2 \leq 2||x-a||^2 + 2||a||^2.
\end{aligned}$$

Note that the last term only depends on x via $||x-a||^2$. So, if we assume that there exist $a \in \mathcal{X}$ and $\beta > 0$ such that $\int \exp(\beta||x-a||^2)\mathrm{d}\nu_p(x) < \infty$, then we can make $\int_{\mathbb{R}^d} \exp\langle\alpha, x\rangle\mathrm{d}\nu_p(x) < \infty$ for all $\alpha \in \mathbb{R}^d$ in the following way. Take any arbitrary $\alpha \in \mathbb{R}^d$, then:

$$\begin{aligned}\int_{\mathbb{R}^d} \exp\langle \alpha, x\rangle \pi(\mathrm{d}x) &\leq \int_{\mathbb{R}^d} \exp\left(\frac{||\alpha||^2}{2\beta} + \beta(||x-a||^2 + ||a||^2)\right)\pi(\mathrm{d}x) \\ &= \exp\left(\frac{||\alpha||^2}{2\beta} + \beta||a||^2)\right) \underbrace{\int_{\mathbb{R}^d} \exp\left(\beta(||x-a||^2)\right)\pi(\mathrm{d}x)}_{<\infty} \\ &< \infty.\end{aligned}$$

In this way, the $T1$ inequality can guarantee that $\int_{\mathbb{R}^d} \exp\langle \alpha, x\rangle \pi(\mathrm{d}x) < \infty$ for all $\alpha \in \mathbb{R}^d$. This is what has been done in Salim et al. 2022 to deduce tightness of the measures for SVGD. Let us now present a theorem that gives the same outcome for the probability measures $(\mu_i)_{i\geq 1}$, but without the need for any condition on the target distribution ν_p.

Theorem 7.15. *Let $(\mu_i)_{i\geq 1}$ be a sequence in $\mathcal{P}(\mathbb{R}^d)$ and consider a target distribution $\nu_p \in \mathcal{P}(\mathbb{R}^d)$. Assume that $\sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p) < \infty$. Then $(\mu_i)_{i\geq 1}$ is both uniformly tight and uniformly integrable in the sense that $\lim_{C\to\infty}\sup_{i\in\mathbb{N}} \int_{\{x\in\mathbb{R}^d \mid ||x||>C\}} ||x|| \mathrm{d}\mu_i(x) = 0$.*

Proof. Note that $\forall i \in \mathbb{N}$ we have that $KL(\mu_i||\nu_p) < \infty$. This means that for all $i \in \mathbb{N}$, μ_i is dominated by ν_p and hence admits a Radon–Nikodym derivative $f_i := \frac{\mathrm{d}\mu_i}{\mathrm{d}\nu_p}$ such that $\int f_i \mathrm{d}\nu_p = 1\ \forall i \in \mathbb{N}$. Observe that we can write the following:

$$\begin{aligned}\int f_i \log f_i \mathrm{d}\nu_p &= \int \log\left(\frac{\mathrm{d}\mu_i}{\mathrm{d}\nu_p}\right)\mathrm{d}\mu_i \\ &= KL(\mu_i||\nu_p).\end{aligned}$$

We will first show uniform integrability and then continue with proving uniform tightness. Let us introduce two inequalities that we are going to use. The first one is that for any nonnegative $a, b \in \mathbb{R}$ and $\sigma \geq 1$: $ab \leq e^{\sigma a} + \frac{1}{\sigma}(b\log b - b + 1)$. This can be derived by noting that $\sup_{a\in\mathbb{R}}\{ab - e^{\sigma a}\} = \frac{b}{\sigma}\left(\log\frac{b}{\sigma} - 1\right) \leq \frac{1}{\sigma}(b\log b - b + 1)$. In the last inequality we used that $-b\log\sigma \leq 0$ for $b \geq 0$ and $\sigma \geq 1$.

We also need the inequality $b\log b - b + 1 > 0$ for all $b > 0$. This can for instance be shown by noting that the function $b \mapsto h(b) := b\log b - b + 1$ is a convex function and that $b = 1$ is a zero of it and at $b = 1$ it is the case that $h'(b) = 0$, so it is the only zero of h by convexity. Now that we have established the necessary inequalities, let us proceed. Take any $C > 0, \sigma \geq 1$ and an arbitrary $i \in \mathbb{N}$:

$$\begin{aligned}
\int_{||x||>C} ||x|| f_i(x) \mathrm{d}\nu_p(x) &\leq \int_{||x||>C} \left(e^{\sigma||x||} + \frac{1}{\sigma}(f_i(x)\log f_i(x) - f_i(x) + 1) \right) \mathrm{d}\nu_p(x) \\
&\leq \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} \int_{\mathcal{X}} \underbrace{(f_i(x)\log f_i(x) - f_i(x) + 1)}_{\geq 0} \mathrm{d}\nu_p(x) \\
&= \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \\
&\quad \frac{1}{\sigma} \Big(\int_{\mathcal{X}} f_i(x)\log f_i(x) \mathrm{d}\nu_p(x) - \underbrace{\int_{\mathcal{X}} f_i(x) \mathrm{d}\nu_p(x)}_{=1} + \underbrace{\int_{\mathcal{X}} \mathrm{d}\nu_p(x)}_{=1} \Big) \\
&= \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} \int_{\mathcal{X}} f_i(x)\log f_i(x) \mathrm{d}\nu_p(x) \\
&= \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} KL(\mu_i||\nu_p). \qquad (91)
\end{aligned}$$

In the first line we have used $ab \leq e^{\sigma a} + \frac{1}{\sigma}(b\log b - b + 1)$ pointwise with $a = ||x||$ and $b = f_i(x)$. In the second line we used that $b\log b - b + 1 > 0$ with $b = f_i(x)$. Note that it is assumed that $\sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p) < \infty$, so for an arbitrary $i \in \mathbb{N}$ we have $\lim_{\sigma\to\infty} \frac{1}{\sigma} KL(\mu_i||\nu_p) = 0$. Because we took $i \in \mathbb{N}$ arbitrary and taking the supremum preserves non-strict inequalities, we have:

$$\begin{aligned}
\sup_{i\in\mathbb{N}} \int_{||x||>C} ||x|| f_i(x) \mathrm{d}\nu_p(x) &\leq \sup_{i\in\mathbb{N}} \left\{ \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} KL(\mu_i||\nu_p) \right\} \\
&= \int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} \sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p).
\end{aligned}$$

We also take the limit of $C \to \infty$ to deduce that:

$$\begin{aligned}
\lim_{C\to\infty} \sup_{i\in\mathbb{N}} \int_{||x||>C} ||x|| f_i(x) \mathrm{d}\nu_p(x) &\leq \lim_{C\to\infty} \left(\int_{||x||>C} e^{\sigma||x||} \mathrm{d}\nu_p(x) + \frac{1}{\sigma} \sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p) \right) \\
&= \frac{1}{\sigma} \sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p).
\end{aligned}$$

Note that the left-hand side of this inequality does not depend on σ, whereas the right-hand side does. Note that $\lim_{\sigma\to\infty} \frac{1}{\sigma} KL(\mu_i||\nu_p) = 0$, as $\sup_{i\in\mathbb{N}} KL(\mu_i||\nu_p) < \infty$. Furthermore, the term on the left-hand side of the inequality is nonnegative. This gives that:

$$\lim_{C\to\infty} \sup_{i\in\mathbb{N}} \int_{||x||>C} ||x|| f_i(x) \mathrm{d}\nu_p(x) = 0.$$

Now note that the f_i are the Radon-Nikodym derivatives and that $\mu_i \ll \nu_p$ for all $i \in \mathbb{N}$. This ultimately gives:

$$\lim_{C\to\infty}\sup_{i\in\mathbb{N}}\int_{||x||>C}||x||\mathrm{d}\mu_i(x)=0.$$

This shows that the sequence of measures $(\mu_i)_{i\geq 1}$ is uniformly integrable.

We will now continue with proving the uniform tightness part. Let us redo the chain of inequalities leading to the inequality in equation (91), but now with $||x||=1$ in the integrand. This gives us for an arbitrary $i\in\mathbb{N}$, arbitrary $C>0$ and arbitrary $\sigma\geq 1$:

$$\int_{||x||>C}f_i(x)\mathrm{d}\nu_p(x)\leq e^{\sigma}\int_{||x||>C}\mathrm{d}\nu_p(x)+\frac{1}{\sigma}KL(\mu_i||\nu_p). \tag{92}$$

This transforms equation (92) into:

$$\mu_i(||X||>C)\leq e^{\sigma}\nu_p(||X||>C)+\frac{1}{\sigma}KL(\mu_i||\nu_p).$$

Taking the limit supremum on both sides:

$$\limsup_{i\to\infty}\mu_i(||X||>C)\leq e^{\sigma}\nu_p(||X||>C)+\frac{1}{\sigma}\limsup_{i\to\infty}KL(\mu_i||\nu_p),$$

Observe that this bound can be made arbitrarily small by choosing σ and C large enough. In other words, for all $\epsilon>0$ we can find constants $C,\sigma<\infty$ such that $\limsup_{i\to\infty}\mu_i(||X||>C)\leq\epsilon$. Hence, the sequence $(\mu_i)_{i\geq 1}$ is uniformly tight.

□

This is a strong result, as it can help us in deducing convergence for SVGD measures in the infinite particle regime. The only condition that is needed is that $\sup_{i\in\mathbb{N}}KL(\mu_i||\nu_p)<\infty$ and this gives that $(\mu_i)_{i\geq 1}$ is uniformly tight and even uniformly integrable.

This is not the complete story, as we want to work towards convergence. The theorem stating a descent result for KL divergence was presented in Theorem 6.6, where the KL-divergence monotonically decreases. However, we could not argue that the KL divergence converges to zero. Let us denote μ_i^{∞} as the SVGD measure at iteration i in the infinite particle regime. We made the assumption that $D(\mu_i^{\infty}||\nu_p)\to 0$ as $i\to\infty \implies \mu_i^{\infty}\Rightarrow\nu_p$ as $i\to\infty$. Under this assumption it is the case that Theorem 6.6 shows that for SVGD iterations it holds true that $\mu_i^{\infty}\Rightarrow\nu_p$ as $i\to\infty$. The problem was that the KSD only metrizes weak convergence under certain assumptions. In Theorem 7.7 it is shown that KSD can detect convergence whenever the measures actually convergence. Detecting non-convergence is also a necessary property and in Theorem 7.5 the conditions are given under which KSD can detect non-convergence. In particular, it shows that KSD can detect non-convergence if the measures are assumed to be tight. So, what we need is that the SVGD measures are forming a uniformly tight sequence to finalise our argument concerning weak convergence of the SVGD measures in the infinite particle regime. This is exactly what has just been shown in Theorem 7.15.

8 Discussion

In this thesis, the general topic of study was Bayesian deep learning. In particular, we were interested in getting insights in the Bayesian paradigm for deep learning. In other words, how can a Bayesian viewpoint help to tackle problems arising in deep learning. That is quite a general question but it correctly mimics our initial broad scope of this thesis, as we started with finding a motivation for Bayesian deep learning.

During this first initial exploration of Bayesian deep learning, it was observed that many methods were coined 'Bayesian', while they are in fact not fully Bayesian in the sense that the posterior does not directly from a prior and a likelihood. Furthermore, in many deep learning papers it was hard to find a proper mathematical structure in which the method was explained. For instance, the deep ensembles paper Lakshminarayanan et al. 2017 contains very little equations, definitions or theorems. In principle that is no problem, but it makes it harder to do mathematical research on a mathematical object (the deep ensemble) that still has to be developed. That was also one of our main tasks in this thesis: finding mathematics in deep learning.

Towards this end, inspiration was found in D'Angelo and Fortuin 2021, where more mathematical structure was used in a (Bayesian) deep learning setting. That paper also hinted at using the concept of a gradient flow in Wasserstein space and SVGD, which are the two main topics of study in this thesis. Uniting SVGD with the concept of Wasserstein gradient flows is particularly hard. For instance, defining a general gradient flow in Section 5.3 already turned out to be quite technical for a general functional on the Wasserstein space. To be able to use that framework, SVGD has to be written as a functional on the Wasserstein space, while we only knew it in terms of a particle update algorithm. This is what we studied in Section 5.8, where by means of a kernel integral operator we could transform the Wasserstein gradient for the KL divergence into a gradient flow corresponding to SVGD. In particular, the discretisation in time of this kernelized Wasserstein gradient flow brought us the SVGD algorithm. This gives a mathematical viewpoint on the SVGD algorithm. Furthermore, it also gives us a way to view the evolution of the measures corresponding to SVGD in Wasserstein space, even though this derivation was slightly informal. In principle, this framework can be used to construct a Wasserstein gradient flow for a functional different from the KL divergence. For instance, we could consider a f-divergence, of which the KL divergence is a concrete example. Then, following the same procedure, we might find a new algorithm that is a discretisation of this f-divergence.

In Section 6 we studied convergence results for SVGD in the infinite particle regime. More precisely, we studied a descent result in Theorem 6.6 for the KL divergence from Liu 2017. To be able to conclude that SVGD could establish weak convergence towards the target distribution, it was necessary to assume that KSD metrizes weak convergence. However, it turned out that this only holds in specific cases, e.g. in the case that a very specific kernel was used. To be able to let the weak convergence result hold in more general scenarios, it would be ideal if we could show that SVGD generates a tight sequence of measures. This would open the door for a more general weak convergence result. This was the topic of the last section of this thesis, Section 7. In this section, we work towards proving uniform tightness of the measures for SVGD under different assumptions. In the end, a theorem is presented that shows that under the mild condition that the KL divergence does not explode along the sequence of measures, it is the case that this sequence is uniformly tight and even uniformly integrable. In turn, under this condition, KSD metrizes weak convergence (as it deals with a tight sequence of measures) and together with the KL divergence descent result it shows that SVGD can weakly converge to the target distribution in the infinite particle regime. However, the KL divergence descent result needs to have sufficiently small step sizes, which are upper bounded by the inverse of the KSD.

Hence, if the KSD is unbounded along the iterations, then the descent result does not hold. It is a-priori hard to check whether this assumption holds and hence has to be assumed. This makes it hard to verify the assumptions of this result in practice. Ideally, we want a KL divergence descent result that has a step-size that is independent of any quantity that is not known at the beginning of running the algorithm.

To conclude, it is good to have one last look at the mountain which is on the cover of this manuscript. After having read this thesis, the Kitzsteinhorn mountain can in an abstract sense be seen as representing a gradient flow problem, where a skier follows a gradient flow along the mountain. This skier might also be the reader of this manuscript, where I sincerely hope that he or she has gained valuable insights into the Bayesian paradigm for deep learning.

Bibliography

Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savare (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media.

Brooks, Steve, Andrew Gelman, Galin Jones, and Xiao-Li Meng (2011). *Handbook of markov chain monte carlo*. CRC press.

Chewi, Sinho, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet (2020). "SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence". In: *Advances in Neural Information Processing Systems* 33, pp. 2098–2109.

D'Angelo, Francesco and Vincent Fortuin (2021). "Repulsive deep ensembles are bayesian". In: *Advances in Neural Information Processing Systems* 34, pp. 3451–3465.

Dupuis, Paul and Richard S Ellis (2011). *A weak convergence approach to the theory of large deviations*. John Wiley & Sons.

Garipov, Timur, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson (2018). "Loss surfaces, mode connectivity, and fast ensembling of dnns". In: *Advances in neural information processing systems* 31.

Gorham, Jackson and Lester Mackey (2017). "Measuring sample quality with kernels". In: *International Conference on Machine Learning*. PMLR, pp. 1292–1301.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hoffmann, Lara and Clemens Elster (2021). "Deep ensembles from a bayesian perspective". In: *arXiv preprint arXiv:2105.13283*.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hunter, John K. (2014). *Notes on Partial Differential Equations*. https://www.math.ucdavis.edu/~hunter/m218a_09/pde_notes.pdf. [Online; accessed 4-April-2023].

Korba, Anna, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin (2021). "Kernel stein discrepancy descent". In: *International Conference on Machine Learning*. PMLR, pp. 5719–5730.

Korba, Anna, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton (2020). "A non-asymptotic analysis for Stein variational gradient descent". In: *Advances in Neural Information Processing Systems* 33, pp. 4672–4682.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Liu, Qiang (2017). "Stein variational gradient descent as gradient flow". In: *Advances in neural information processing systems* 30.

Liu, Qiang, Jason Lee, and Michael Jordan (2016). "A kernelized Stein discrepancy for goodness-of-fit tests". In: *International conference on machine learning*. PMLR, pp. 276–284.

Liu, Qiang and Dilin Wang (2016). "Stein variational gradient descent: A general purpose bayesian inference algorithm". In: *Advances in neural information processing systems* 29.

Lu, Jianfeng, Yulong Lu, and James Nolen (2019). "Scaling limit of the Stein variational gradient descent: The mean field regime". In: *SIAM Journal on Mathematical Analysis* 51.2, pp. 648–671.

Rasmussen, Carl and Zoubin Ghahramani (2000). "Occam's razor". In: *Advances in neural information processing systems* 13.

Remes, Sami, Markus Heinonen, and Samuel Kaski (2017). "Non-stationary spectral kernels". In: *Advances in neural information processing systems* 30.

Salim, Adil, Lukang Sun, and Peter Richtarik (2022). "A convergence theory for SVGD in the population limit under Talagrand's inequality T1". In: *International Conference on Machine Learning*. PMLR, pp. 19139–19152.

Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

Vaart, AW van der and Jon A Wellner (2023). "Empirical processes". In: *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, pp. 127–384.

Villani, Cédric et al. (2009). *Optimal transport: old and new*. Vol. 338. Springer.

Wang, Ziyu, Tongzheng Ren, Jun Zhu, and Bo Zhang (2019). "Function space particle optimization for bayesian neural networks". In: *arXiv preprint arXiv:1902.09754*.

Wasserman, Larry (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Williams, Christopher KI and Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA.

Wilson, Andrew G and Pavel Izmailov (2020). "Bayesian deep learning and a probabilistic perspective of generalization". In: *Advances in neural information processing systems* 33, pp. 4697–4708.

Wilson, Andrew Gordon (2020). "The case for Bayesian deep learning". In: *arXiv preprint arXiv:2001.10995*.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2021). "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3, pp. 107–115.

Zhou, Ding-Xuan (2008). "Derivative reproducing properties for kernel methods in learning theory". In: *Journal of computational and Applied Mathematics* 220.1-2, pp. 456–463.

A Proofs

A.1 Proof of Theorem 3.7

Proof. This proof is from Liu, Lee, et al. 2016, but this version is more elaborate. Let us introduce a short-hand notation for $\boldsymbol{\phi}^*_{q,p}$, which was the function $x' \mapsto \boldsymbol{\phi}^*_{q,p}(x') = \mathrm{E}_{X\sim q}[\mathcal{A}_p k(x', X)]$. We will use the short-hand notation $\boldsymbol{\beta} := \boldsymbol{\phi}^*_{q,p}$ Let us start by proving that $S(p,q) = ||\boldsymbol{\beta}||^2_{\mathcal{H}^d}$. Let us start with a preliminary definition (from Liu, Lee, et al. 2016) of KSD that is more 'fundamental' than the definition we gave in Definition 3.6. This can also be skipped if the reader wants to stick to the definition of KSD as in Definition 3.6, then continue reading with equation (93).

Definition A.1. The kernelized Stein discrepancy (KSD) between densities p and q is defined as:

$$S(q,p) = \mathrm{E}_{x,x'\sim q}[\boldsymbol{\delta}_{p,q}(x)^T k(x,x')\boldsymbol{\delta}_{p,q}(x')],$$

where $\boldsymbol{\delta}_{p,q}(x) = \boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)$ is the difference between score functions of p and q and x, x' are i.i.d. draws from the density q.

Using this definition and the reproducing property of kernels in the RKHS: $k(x,x') = \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}$:

$$\begin{aligned}
S(p,q) &= \mathrm{E}_{x,x'\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^T k(x,x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))] \\
&= \mathrm{E}_{x,x'\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^T \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))] \\
&= \mathrm{E}_{x\sim q}\,\mathrm{E}_{x'\sim q}[\sum_{\ell=1}^{d}(\boldsymbol{s}_p^\ell(x) - \boldsymbol{s}_q^\ell(x))(\boldsymbol{s}_p^\ell(x') - \boldsymbol{s}_q^\ell(x'))\langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}] \\
&= \sum_{\ell=1}^{d}\mathrm{E}_{x\sim q}[(\boldsymbol{s}_p^\ell(x) - \boldsymbol{s}_q^\ell(x))\,\mathrm{E}_{x'\sim q}[(\boldsymbol{s}_p^\ell(x') - \boldsymbol{s}_q^\ell(x'))\langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}]] \\
&= \sum_{\ell=1}^{d}\mathrm{E}_{x\sim q}[(\boldsymbol{s}_p^\ell(x) - \boldsymbol{s}_q^\ell(x))\langle k(x,\cdot), \mathrm{E}_{x'\sim q}[(\boldsymbol{s}_p^\ell(x') - \boldsymbol{s}_q^\ell(x'))k(x',\cdot)]\rangle_{\mathcal{H}}] \\
&= \sum_{\ell=1}^{d}\left\langle \mathrm{E}_{x\sim q}[(\boldsymbol{s}_p^\ell(x) - \boldsymbol{s}_q^\ell(x))k(x,\cdot)], \mathrm{E}_{x'\sim q}[k(x',\cdot)(\boldsymbol{s}_p^\ell(x') - \boldsymbol{s}_q^\ell(x'))]\right\rangle_{\mathcal{H}} \\
&= \sum_{\ell=1}^{d}\langle \boldsymbol{\beta}_\ell, \boldsymbol{\beta}_\ell\rangle_{\mathcal{H}} \\
&= ||\boldsymbol{\beta}||^2_{\mathcal{H}^d}.
\end{aligned}$$

Here we used Lemma 2.3 from Liu, Lee, et al. 2016, i.e. $\mathrm{E}_{x\sim q}[\mathcal{A}_p k_{x'}(x)] = \mathrm{E}_{x\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))k_{x'}(x)]$, as $k_{x'}(\cdot) := k(x',\cdot)$ is in the Stein class of q.

Let us now make the connection with the definition that we used for KSD. For any $\boldsymbol{\phi} \in \mathcal{H}^d$ we have:

$$\begin{aligned}\langle \boldsymbol{\phi}, \boldsymbol{\beta} \rangle_{\mathcal{H}^d} &= \sum_{\ell=1}^{d} \langle \phi_\ell, \mathrm{E}_{x\sim q}[\boldsymbol{s}_p^\ell(x)k(x,\cdot) + \nabla_{x_\ell} k(x,\cdot)] \rangle_{\mathcal{H}} \\ &= \sum_{\ell=1}^{d} \mathrm{E}_{x\sim q}[\boldsymbol{s}_p^\ell(x) \langle \phi_\ell, k(x,\cdot)\rangle_{\mathcal{H}} + \langle \phi_\ell, \nabla_{x_l} k(x,\cdot)\rangle_{\mathcal{H}}] \\ &= \sum_{\ell=1}^{d} \mathrm{E}_{x\sim q}[\boldsymbol{s}_p^\ell(x)\phi_\ell(x) + \nabla_{x_\ell}\phi_\ell(x)] \\ &= \mathrm{E}_{x\sim q}[\mathrm{tr}(\mathcal{A}_p \boldsymbol{\phi}(x))],\end{aligned} \tag{93}$$

where we have used the reproducing property of the kernel and the non-trivial fact that $\nabla_x \phi(x) = \langle \phi, \nabla_x k(x,\cdot)\rangle_{\mathcal{H}}$ from Zhou 2008.

We will now work towards the equality $S(p,q) = \max_{\boldsymbol{\phi}\in\mathcal{H}^d} \left\{ \mathrm{E}_q[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi})]^2 \mid ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1 \right\}$. To this end, we will first establish the following equality $||\boldsymbol{\beta}||_{\mathcal{H}^d} = \max_{\boldsymbol{\phi}\in\mathcal{H}^d} \{\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle \mid ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\}$. By Cauchy-Schwarz we have the following inequality:

$$\begin{aligned}|\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d}| &\leq ||\boldsymbol{\phi}||_{\mathcal{H}^d} ||\boldsymbol{\beta}||_{\mathcal{H}^d} \\ &\leq ||\boldsymbol{\beta}||_{\mathcal{H}^d}, \quad \forall \boldsymbol{\phi} \text{ s.t. } ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1.\end{aligned}$$

Hence, we have that $\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d} \leq ||\boldsymbol{\beta}||_{\mathcal{H}^d} \quad \forall \boldsymbol{\phi}$ such that $||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1$. Let us pick $\boldsymbol{\phi} = \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_{\mathcal{H}^d}}$. Observe that this function has norm equal to 1. Furthermore, it satisfies the following equality:

$$\begin{aligned}\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d} &= \left\langle \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_{\mathcal{H}^d}}, \boldsymbol{\beta} \right\rangle_{\mathcal{H}^d} \\ &= \frac{1}{||\boldsymbol{\beta}||_{\mathcal{H}^d}} \langle \boldsymbol{\beta}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d} \\ &= ||\boldsymbol{\beta}||_{\mathcal{H}^d}.\end{aligned}$$

So we have now shown that a maximum is bounded and attained and hence we can state thate that $||\boldsymbol{\beta}||_{\mathcal{H}^d} = \max_{\boldsymbol{\phi}\in\mathcal{H}^d} \{\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle \mid ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\}$. By the previously established equality $\langle \boldsymbol{\phi}, \boldsymbol{\beta}\rangle = \mathrm{E}_{x\sim q}[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi}(x))]$ and $S(p,q) = ||\boldsymbol{\beta}||^2_{\mathcal{H}^d}$ the result follows.

Also observe that $\boldsymbol{\beta}(\cdot) = \mathrm{E}_{x\sim q}[\boldsymbol{s}_p(x)k(x,\cdot) + \nabla_x k(x,\cdot)] = \int_{x\in\mathcal{X}} (\boldsymbol{s}_p(x)k(x,\cdot) + \nabla_x k(x,\cdot))q(x)\mathrm{d}x$. We know that $(x,x') \mapsto k(x,x')$ is in the Stein class of q, i.e. it is smooth and satisfies equation (11) for any fixed x':

$$\int_{x\in\mathcal{X}} \nabla_x (k(x,x')q(x))\mathrm{d}x = 0, \quad \text{and} \int_{x\in\mathcal{X}} \nabla_x (k(x',x)q(x))\mathrm{d}x = 0.$$

We now use use the second equality to show that $\nabla_x k(x,\cdot)$ is also in the Stein class of q. Take any arbitrary fixed x' and then we have:

$$\int_{x\in\mathcal{X}} \nabla_x(q(x)\nabla_{x'}k(x',x))\mathrm{d}x = \nabla_{x'}\int_{x\in\mathcal{X}} \nabla_x(q(x)k(x',x))\mathrm{d}x = 0.$$

This shows that $\nabla_x k(x,\cdot)$ is in the Stein class of q. Let us now show that $\boldsymbol{\beta}$ is in the Stein class of q, using that we know that $k(x,\cdot)$ and $\nabla_x k(x,\cdot)$ are in the Stein class of q:

$$\int_{x\in\mathcal{X}} \nabla_x(q(x)\,\mathrm{E}_{x'\sim q}[\boldsymbol{s}_p(x')k(x',x) + \nabla_{x'}k(x',x)])\mathrm{d}x =$$
$$\mathrm{E}_{x'\sim q}\left[\boldsymbol{s}_p(x')\underbrace{\int_{x\in\mathcal{X}} \nabla_x(q(x)k(x',x))\mathrm{d}x}_{=0\ \forall x'\in\mathcal{X}} + \underbrace{\int_{x\in\mathcal{X}} \nabla_x(q(x)\nabla_{x'}k(x',x))\mathrm{d}x}_{=0\ \forall x'\in\mathcal{X}}\right] = 0$$

□

A.2 Proof of Theorem 3.8

We will prove Theorem 3.8, following Liu and D. Wang 2016. To this end, we will first state and prove a preliminary result in the form of a lemma that the authors also use.

Lemma A.2. *Let q and p be two smooth densities (continuously differentiable) and $\boldsymbol{T} = \boldsymbol{T}_\epsilon(x)$, $\boldsymbol{T}_\epsilon : \mathcal{X} \to \mathcal{X}$ a bijective transform on $\mathcal{X}$ that is indexed by ϵ. Assume $\boldsymbol{T}$ is differentiable with respect to x and ϵ. Also assume that $\boldsymbol{T}^{-1}$ is differentiable and its Jacobian is nonsingular in its domain. Define $q_{[\boldsymbol{T}]}$ as the density of Z, with $Z = \boldsymbol{T}_\epsilon(X)$, where X has density q. We denote the score function of a density p as $\boldsymbol{s}_p = \nabla \log p$. Then we have:*

$$\nabla_\epsilon KL(q_{[\boldsymbol{T}]}||p) = -\,\mathrm{E}_{x\sim q}[\boldsymbol{s}_p(\boldsymbol{T}(x))^T\nabla_\epsilon\boldsymbol{T}(x) + \mathrm{tr}(\nabla_x\boldsymbol{T}(x))^{-1}\cdot\nabla_\epsilon\nabla_x\boldsymbol{T}(x))].$$

Proof. Let us denote $q_{[\boldsymbol{T}^{-1}]}$ as the density of $Z = \boldsymbol{T}^{-1}(X)$ when $X \sim q$, then by the change of variables formula:

$$q_{[\boldsymbol{T}^{-1}]}(z) = q(\boldsymbol{T}(z))\cdot|\det(\nabla_z\boldsymbol{T}(z))|.$$

By this change of variables, we also have $KL(q_{[\boldsymbol{T}]}||p) = KL(q||p_{[\boldsymbol{T}^{-1}]})$. This is due to the fact that KL divergence is invariant under parameter transformations:

$$
\begin{aligned}
KL(q_{[\boldsymbol{T}]}||p) &= \int_{\mathcal{X}} \log\left(\frac{q_{[\boldsymbol{T}]}(x)}{p(x)}\right) q_{[\boldsymbol{T}]}(x)\mathrm{d}x \\
&= \int_{\mathcal{X}} \log\left(\frac{q(\boldsymbol{T}^{-1}(x)) \cdot |\det(\nabla \boldsymbol{T}^{-1}(x))|}{p(x)}\right) q(\boldsymbol{T}^{-1}(x)) \cdot |\det(\nabla \boldsymbol{T}^{-1}(x))|\mathrm{d}x \\
&= \int_{\mathcal{X}} \log\left(\frac{q(y)|\det(\nabla \boldsymbol{T}^{-1}(\boldsymbol{T}(y)))|}{p(\boldsymbol{T}(y))}\right) q(y)\mathrm{d}y, \quad \text{for } y = \boldsymbol{T}^{-1}(x), \\
&= \int_{\mathcal{X}} \log\left(\frac{q(y)}{p(\boldsymbol{T}(y))|\det(\nabla \boldsymbol{T}(y))|}\right) q(y)\mathrm{d}y \\
&= \int_{\mathcal{X}} \log\left(\frac{q(y)}{p_{[\boldsymbol{T}^{-1}]}(y)}\right) q(y)\mathrm{d}y \\
&= KL(q||p_{[\boldsymbol{T}^{-1}]}),
\end{aligned}
$$

where we have used a change of variables in the third equality with the introduction of a new variable y. In the fourth equality we have used a result of the inverse function theorem that the matrix inverse of the Jacobian for an invertible function is the Jacobian matrix of the inverse function, i.e. $(\nabla \boldsymbol{T}(x))^{-1} = \nabla \boldsymbol{T}^{-1}(\boldsymbol{T}(x))$. We also used the fact that the determinant of a matrix is equal to the the reciprocal of the determinant of the inverse of that matrix ($|A| = 1/|A^{-1}|$).

This also yields the following equality:

$$
\begin{aligned}
\nabla_\epsilon KL(q_{[\boldsymbol{T}]}||p) &= \nabla_\epsilon KL(q||p_{[\boldsymbol{T}^{-1}]}) \\
&= \nabla_\epsilon \int_{\mathcal{X}} \log\left(\frac{q(x)}{p_{[\boldsymbol{T}^{-1}]}(x)}\right) q(x)\mathrm{d}x \\
&= \int_{\mathcal{X}} \nabla_\epsilon(\log q(x) - \log p_{[\boldsymbol{T}^{-1}]}(x)) q(x)\mathrm{d}x \\
&= \int_{\mathcal{X}} -\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(x) q(x)\mathrm{d}x \\
&= -\operatorname{E}_{X\sim q}[\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(X)].
\end{aligned}
$$

Let us now calculate $\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(x)$:

$$
\begin{aligned}
\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(x) &= \frac{1}{p(\boldsymbol{T}(x))|\det(\nabla_x \boldsymbol{T}(x))|} \nabla_\epsilon p_{[\boldsymbol{T}^{-1}]}(x) \\
&= \frac{1}{p(\boldsymbol{T}(x))|\det(\nabla_x \boldsymbol{T}(x))|} (\nabla_\epsilon p(\boldsymbol{T}(x))|\det(\nabla_x \boldsymbol{T}(x))|) \\
&= \frac{1}{p(\boldsymbol{T}(x))|\det(\nabla_x \boldsymbol{T}(x))|} (|\det(\nabla_x \boldsymbol{T}(x))|\nabla_\epsilon p(\boldsymbol{T}(x)) + p(\boldsymbol{T}(x))\nabla_\epsilon |\det(\nabla_x \boldsymbol{T}(x))|).
\end{aligned}
$$

Let us observe now that the derivative with respect to a scalar of a scalar valued function with vector input is $\frac{\partial g(\boldsymbol{u})}{\partial \epsilon} = (\nabla_{\boldsymbol{u}} g(\boldsymbol{u}))^T \frac{\partial \boldsymbol{u}}{\partial \epsilon}$, with $\boldsymbol{u} = \boldsymbol{u}(\epsilon)$, for ϵ being a scalar on which $\boldsymbol{u}$ depends.

Following this fact, we have $\nabla_\epsilon p(\boldsymbol{T}(x)) = (\nabla_x p(\boldsymbol{T}(x)))^T \nabla_\epsilon \boldsymbol{T}(x)$. By Jacobi's formula we also have the following identity:

$$\nabla_\epsilon \det(\nabla_x \boldsymbol{T}(x)) = \det(\nabla_x \boldsymbol{T}(x)) \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right).$$

Therefore, we have:

$$\begin{aligned}\nabla_\epsilon \left|\det(\nabla_x \boldsymbol{T}(x))\right| &= \frac{\det(\nabla_x \boldsymbol{T}(x))}{\left|\det(\nabla_x \boldsymbol{T}(x))\right|} \det(\nabla_x \boldsymbol{T}(x)) \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right) \\ &= \left|\det(\nabla_x \boldsymbol{T}(x))\right| \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right).\end{aligned}$$

Continuing with the derivation of $\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(x)$:

$$\begin{aligned}\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(x) &= \frac{1}{p(\boldsymbol{T}(x))\left|\det(\nabla_x \boldsymbol{T}(x))\right|} \Big(\left|\det(\nabla_x \boldsymbol{T}(x))\right| (\nabla_x p(\boldsymbol{T}(x)))^T \nabla_\epsilon \boldsymbol{T}(x) + \\ &\quad p(\boldsymbol{T}(x))\left|\det(\nabla_x \boldsymbol{T}(x))\right| \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right) \\ &= \frac{1}{p(\boldsymbol{T}(x))} (\nabla_x p(\boldsymbol{T}(x)))^T \nabla_\epsilon \boldsymbol{T}(x) + \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right) \\ &= \left(\boldsymbol{s}_p(\boldsymbol{T}(x))\right)^T \nabla_\epsilon \boldsymbol{T}(x) + \operatorname{tr}\left((\nabla_x \boldsymbol{T}(x))^{-1} \nabla_\epsilon \nabla_x \boldsymbol{T}(x)\right)\end{aligned}$$

□

Now that we have proven this lemma, it is time to prove the theorem:

Proof. In the assumptions of the theorem we have $\boldsymbol{T}(x) = x + \epsilon \boldsymbol{\phi}(x)$. This gives that for $\epsilon = 0$:

$$\boldsymbol{T}(x) = x, \text{ and } \nabla_x \boldsymbol{T}(x) = I.$$

We also have that for $\boldsymbol{T}(x) = x + \epsilon \boldsymbol{\phi}(x)$: $\nabla_\epsilon \boldsymbol{T}(x) = \boldsymbol{\phi}(x)$ and $\nabla_\epsilon \nabla_x \boldsymbol{T}(x) = \nabla_x \boldsymbol{\phi}(x)$.

Hence, using the result of the lemma and evaluating it at $\epsilon = 0$ and filling the terms from above in yields:

$$\begin{aligned}\nabla_\epsilon KL(q_{[\boldsymbol{T}]} || p)\Big|_{\epsilon=0} &= -\operatorname{E}_q[\boldsymbol{s}_p(\boldsymbol{T}(X))^T \nabla_\epsilon \boldsymbol{T}(X) + \operatorname{tr}(\nabla \boldsymbol{T}(X))^{-1} \cdot \nabla_\epsilon \nabla \boldsymbol{T}(X))] \\ &= -\operatorname{E}_q[\boldsymbol{s}_p(X)^T \boldsymbol{\phi}(X) + \operatorname{tr}((I)^{-1} \cdot \nabla \boldsymbol{\phi}(X))] \\ &= -\operatorname{E}_q[\boldsymbol{s}_p(X)^T \boldsymbol{\phi}(x) + \operatorname{tr}(\nabla \boldsymbol{\phi}(X))] \\ &= -\operatorname{E}_q[\operatorname{tr}(\boldsymbol{s}_p(X)(\boldsymbol{\phi}(X))^T) + \operatorname{tr}(\nabla \boldsymbol{\phi}(X))] \\ &= -\operatorname{E}_q[\operatorname{tr}((\nabla \log p(X))(\boldsymbol{\phi}(X))^T) + \operatorname{tr}(\nabla \boldsymbol{\phi}(X))] \\ &= -\operatorname{E}_q[\operatorname{tr}((\nabla \log p(X))(\boldsymbol{\phi}(X))^T + \nabla \boldsymbol{\phi}(X))] \\ &= -\operatorname{E}_q[\operatorname{tr}(\mathcal{A}_p \boldsymbol{\phi}(X))],\end{aligned}$$

with $\mathcal{A}_p \boldsymbol{\phi}(x) = (\nabla_x \log p(x))(\boldsymbol{\phi}(x))^T + \nabla \boldsymbol{\phi}(x)$. □

A.3 Proof of Theorem 6.6

We will follow Liu 2017 for this proof and slightly elaborate on it.

Proof. Throughout this proof, we let $\mathcal{S}_p$ denote the operator $\boldsymbol{\phi} \mapsto \mathcal{S}_p\boldsymbol{\phi} := \boldsymbol{s}_p^T\boldsymbol{\phi} + \nabla \cdot \boldsymbol{\phi}$, for vector valued functions $\boldsymbol{\phi} = [\phi_1, \dots, \phi_d]^T$. Note that this is equal to $\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi})$, for the vector valued $\boldsymbol{\phi} = [\phi_1, \dots, \phi_d]^T$, where $\mathcal{A}_p$ is defined in Definition 3.1.
1. If Stein's identity, for the density p, is satisfied for all $\boldsymbol{\phi} \in \mathcal{H}^d$, then $\boldsymbol{\phi}^*_{\nu_p,p}$ is the zero map and ν_p is a fixed point of the pushforward measure Φ_p.

2. Let us use as notation $\mu_l = \mu_l^\infty$ in what follows. We can rewrite the following term:

$$\begin{aligned}
KL(\mu_{l+1}||\nu_p) - KL(\mu_l||\nu_p) &= KL((\boldsymbol{T}_{\mu_l,p})_\#\mu_l||\nu_p) - KL(\mu_l||\nu_p) \\
&= KL(\mu_l||(\boldsymbol{T}^{-1}_{\mu_l,p})_\#\nu_p) - KL(\mu_l||\nu_p), \quad \text{by lemma A.2 in Liu 2017,} \\
&= -\,\mathrm{E}_{x\sim\mu_l}[\log p(\boldsymbol{T}_{\mu_l,p}(x)) + \log\det(\nabla\boldsymbol{T}_{\mu_l,p}(x)) - \log p(x)] \\
&= \mathrm{E}_{x\sim\mu_l}[\log p(x) - \log p(\boldsymbol{T}_{\mu_l,p}(x)) - \log\det(\nabla\boldsymbol{T}_{\mu_l,p}(x))], \qquad (94)
\end{aligned}$$

where in the last equality we have used that if the measure μ has density q, then the density of q' of the measure $\mu' = \Phi(\mu)$ ($\Phi_p : \mu \mapsto (\boldsymbol{T}_{\mu,p})_\#\mu$) is given by $z \mapsto q'(z) = q(\boldsymbol{T}^{-1}_{\mu,p}(z)) \cdot |\det(\nabla\boldsymbol{T}^{-1}_{\mu,p}(z))|$. This way we can also get the density of $(\boldsymbol{T}^{-1}_{\mu_l,p})_\#\nu_p$ in an analogous way, as every inverse of $\boldsymbol{T}$ should become simply $\boldsymbol{T}$ and vice versa. Let us remind ourselves that $\boldsymbol{T}_{\mu_l,p} : x \mapsto T_{\mu_l,p}(x) = x + \epsilon\boldsymbol{\phi}^*_{\mu_l,p}(x)$ and let us define $x_s = x + \epsilon s\boldsymbol{\phi}^*_{\mu_l,p}$, $\forall s \in [0,1]$. We can bound the term $\log p(x) - \log p(\boldsymbol{T}_{\mu_l,p}(x))$ as follows,

$$\begin{aligned}
\log p(x) - \log p(\boldsymbol{T}_{\mu_l,p}(x)) &= \int_0^1 \nabla_s \log p(x_s)\mathrm{d}s, \quad \text{by the Fundamental Theorem of Calculus,} \\
&= -\int_0^1 \left(\nabla_x \log p(x)^T\Big|_{x=x_s}\right)(\epsilon\boldsymbol{\phi}^*_{\mu_l,p}(x))\mathrm{d}s, \quad \text{by the chain rule,} \\
&= -\epsilon\nabla_x \log p(x)^T\boldsymbol{\phi}^*_{\mu_l,p}(x) \\
&\quad - \int_0^1 (\nabla_x \log p(x)^T|_{x=x_s} - \nabla_x \log p(x))^T(\epsilon\boldsymbol{\phi}^*_{\mu_l,p}(x))\mathrm{d}s, \text{ adding zero,} \\
&\leq -\epsilon\nabla_x \log p(x)^T\boldsymbol{\phi}^*_{\mu_l,p}(x) + \epsilon^2||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2^2 \int_0^1 s\mathrm{d}s, \text{ by CS,} \\
&= -\epsilon\nabla_x \log p(x)^T\boldsymbol{\phi}^*_{\mu_l,p}(x) + \frac{\epsilon^2}{2}||\nabla\log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2^2. \qquad (95)
\end{aligned}$$

In the first equality we have used the fundamental theorem of calculus (FTC) which obviously holds if p is differentiable, but also if p is only assumed to be weakly differentiable. By Theorem 3.60 in Hunter 2014 the fundamental theorem of calculus also holds for these type of functions. In the fourth line we have used the Cauchy Schwarz (CS) inequality as $|\nabla_x \log p(x_s) - \nabla_x \log p(x)|^T\boldsymbol{\phi}^*_{\mu_l,p}(x) \leq ||\nabla_x \log p(x_s) - \nabla_x \log p(x)||_2 \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2$. Furthermore, we have used that $||x_s - x|| = \epsilon s||\boldsymbol{\phi}^*_{\mu_l,p}(x)||$ to get:

$$|\nabla_x \log p(x_s) - \nabla_x \log p(x)|^T \boldsymbol{\phi}^*_{\mu_l,p}(x) \leq \sup_{x \neq x_s} \left\{ \frac{||\nabla_x \log p(x_s) - \nabla_x \log p(x)||_2}{||x_s - x||_2} \right\} ||x_s - x||_2 ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2. \tag{96}$$

In fact, we have used the 2-Lip norm, defined as $||f||_{\text{Lip}} := \sup_{x \neq y} \left\{ \frac{||f(x)-f(y)||_2}{||x-y||_2} \right\}$ for $f : \mathbb{R}^n \to \mathbb{R}^m$.

Let us now present a useful lemma to bound $\log |\det(\nabla \boldsymbol{T}_{\mu_l,p}(x))|$:

Lemma A.3 (Lemma A.1 from Liu 2017)**.** *Let B be a square matrix and denote by $||B||_F = \sqrt{\sum_{ij} b_{ij}^2}$ its Frobenius norm. Let ϵ be a positive number satisfying $0 \leq \epsilon < \frac{1}{\rho(B+B^T)}$, with $\rho(\cdot)$ denoting the spectral radius of a matrix. Then it is the case that $I + \epsilon(B + B^T)$ is positive definite and*

$$\log |\det(I + \epsilon B)| \geq \epsilon \operatorname{tr}(B) - \epsilon^2 \frac{||B||_F^2}{1 - \epsilon\rho(B + B^T)}.$$

Using an even smaller ϵ, i.e. $0 \leq \epsilon \leq \frac{1}{2\rho(B+B^T)}$ we get:

$$\log |\det(I + \epsilon B)| \geq \epsilon \operatorname{tr}(B) - 2\epsilon^2 ||B||_F^2.$$

Proof. See Liu 2017 for a proof. □

In this lemma, take $B = \nabla \boldsymbol{\phi}^*_{\mu_l,p}$ and the second bound in the lemma for the smallest ϵ, i.e. $\epsilon \leq \frac{1}{2\rho(B+B^T)}$. This yields:

$$\begin{aligned} \log |\det(\underbrace{\nabla \boldsymbol{T}_{\mu_l,p}(x)}_{I+\epsilon B})| &\geq \epsilon \operatorname{tr}(\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)) - 2\epsilon^2 ||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F^2 \\ &= \epsilon \nabla \cdot \boldsymbol{\phi}^*_{\mu_l,p}(x) - 2\epsilon^2 ||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F^2. \end{aligned} \tag{97}$$

Combing the bounds in equations (95),(97) to bound (94) yields:

$$\begin{aligned} KL(\mu_{l+1}) - KL(\mu_l || \nu_p) &\leq -\epsilon \, \mathrm{E}_{\mu_l}[\mathcal{S}_p \boldsymbol{\phi}^*_{\mu_l,p}] + \Delta \\ &= \epsilon D(\mu_l || \nu_p)^2 + \Delta, \end{aligned} \tag{98}$$

with $\Delta = \epsilon^2 \, \mathrm{E}_{x \sim \mu_l}[\frac{1}{2} ||\nabla \log p||_{\text{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2^2 + 2||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F^2]$. We can try to bound the terms $||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2$ and $||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F$. We can take advantage of the fact that we are working in a RKHS and hence the reproducing property of the RKHS can be exploited. Let us write $\boldsymbol{\phi}^*_{\mu_l,p} = [\phi_1, \ldots, \phi_d]^T$ for $\phi_i \in \mathcal{H} \; \forall i = 1, \ldots, d$, with $\boldsymbol{\phi}^*_{\mu_l,p} \in \mathcal{H}^d = \underbrace{\mathcal{H} \times \cdots \times \mathcal{H}}_{d \text{ times}}$. We then have that

$$\phi_i(x) = \langle \phi_i(\cdot), k(x, \cdot) \rangle_{\mathcal{H}_0}, \quad \partial_{x_j} \phi_i(x) = \langle \phi_i(\cdot), \partial_{x_j} k(x, \cdot) \rangle_{\mathcal{H}_0} \quad \forall i, j = 1, \ldots, d.$$

We also use that $||\boldsymbol{\phi}^*_{\mu_l,p}||^2_{\mathcal{H}^d} = \sum_{i=1}^d ||\phi_i||^2_{\mathcal{H}} = D(\mu_l||\nu_p)^2$ by a result from Theorem 3.7. This gives the following chain of (in)equalities:

$$\begin{aligned}
||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2^2 &= \sum_{i=1}^d \phi_i(x)^2 \\
&= \sum_{i=1}^d (\langle k(x,\cdot), \phi_i(\cdot)\rangle_{\mathcal{H}})^2, \quad \text{reproducing property,} \\
&\leq \sum_i ||k(x,\cdot)||^2_{\mathcal{H}} \cdot ||\phi_i||^2_{\mathcal{H}}, \quad \text{Cauchy Schwarz inequality,} \\
&= k(x,x) \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}||^2_{\mathcal{H}^d}, \quad \text{reproducing property,} \\
&= k(x,x) D(\mu_l||\nu_p)^2.
\end{aligned}$$

Let us also give a bound for $||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F$:

$$\begin{aligned}
||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_F^2 &= \sum_{i,j=1}^d \partial_{x_j}\phi_i(x)^2 \\
&= \sum_{i,j=1}^d (\langle \partial_{x_j} k(x,\cdot), \phi_i(\cdot)\rangle_{\mathcal{H}})^2, \quad \text{reproducing property,} \\
&\leq \sum_{i,j=1}^d ||\partial_{x_j} k(x,\cdot)||^2_{\mathcal{H}} \cdot ||\phi_i||^2_{\mathcal{H}}, \quad \text{Cauchy Schwarz inequality,} \\
&= \sum_{i,j=1}^d \partial_{x_j}\partial_{x'_j} k(x,x')|_{x=x'} \cdot ||\phi_i||^2_{\mathcal{H}} \\
&= \sum_{i=1}^d ||\phi_i||^2_{\mathcal{H}} \sum_{j=1}^d \partial_{x_j}\partial_{x'_j} k(x,x')|_{x'=x} \\
&= \nabla_{xx'} k(x,x) \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}||^2_{\mathcal{H}^d} \\
&= \nabla_{xx'} k(x,x) D(\mu_l||\nu_p)^2. \qquad (99)
\end{aligned}$$

In the fourth line we have used that $||\partial_{x_j} k(x,\cdot)||^2_{\mathcal{H}} = \langle \partial_{x_j} k(x,\cdot), \partial_{x_j} k(x,\cdot)\rangle_{\mathcal{H}} = \partial_{x'_j}\partial_{x_j} k(x,x')|_{x'=x}$.

Now that we have two bounds available for $||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2$ and $||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F$. We can now try to bound Δ:

$$
\begin{aligned}
\Delta &= \epsilon^2 \, \mathrm{E}_{x\sim\mu_l}\left[\frac{1}{2}||\nabla \log p||_{\mathrm{Lip}} \cdot ||\boldsymbol{\phi}^*_{\mu_l,p}(x)||_2^2 + 2||\nabla \boldsymbol{\phi}^*_{\mu_l,p}(x)||_F^2\right] \\
&\leq \epsilon^2 \, \mathrm{E}_{x\sim\mu_l}\left[\frac{1}{2}||\nabla \log p||_{\mathrm{Lip}} \cdot k(x,x) D(\mu_l||\nu_p)^2 + 2\nabla_{xx'}k(x,x) D(\mu_l||\nu_p)^2\right] \\
&= \epsilon^2 D(\mu_l||\nu_p)^2 \, \mathrm{E}_{x\sim\mu_l}\left[\frac{1}{2}||\nabla \log p||_{\mathrm{Lip}} k(x,x) + 2\nabla_{xx'}k(x,x)\right] \\
&= \epsilon^2 D(\mu_l||\nu_p)^2 R.
\end{aligned}
$$

Combining this with equation (98) yields:

$$
KL(\mu_{l+1}||\nu_p) - KL(\mu_l||\nu_p) \leq -\epsilon(1-\epsilon R) D(\mu_l||\nu_p)^2.
$$

□

A.4 Proof of Proposition 5.28

Proof. We follow Korba, Salim, et al. 2020.

Let us evaluate the time derivative of the KL divergence:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} KL(\mu_t||\pi) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \log\left(\frac{\mu_t}{\pi}\right) \mathrm{d}\mu_t \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\int \log\left(\frac{\rho_t}{p}\right) \rho_t \mathrm{d}x \\
&= \int \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\frac{\rho_t}{p}\right) \rho_t \mathrm{d}x \\
&= \int \left(\log\left(\frac{\rho_t}{p}\right)\frac{\partial}{\partial t}\rho_t + \rho_t \frac{\partial}{\partial t}\log\left(\frac{\rho_t}{p}\right)\right) \mathrm{d}x \\
&= \int \left(\log\left(\frac{\rho_t}{p}\right)\frac{\partial}{\partial t}\rho_t + \frac{\partial}{\partial t}\rho_t\right) \mathrm{d}x \\
&= \int \log\left(\frac{\rho_t}{p}\right)\frac{\partial}{\partial t}\rho_t \mathrm{d}x + \frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\int \rho_t \mathrm{d}x}_{=1} \\
&= \int \log\left(\frac{\rho_t}{p}\right)\frac{\partial}{\partial t}\rho_t \mathrm{d}x.
\end{aligned}
$$

We now remind ourselves that $(\mu_t)_{t\geq 0}$ satisfies a continuity equation:

$$
\frac{\partial}{\partial}\mu_t = -\nabla \cdot (\mu_t v_t).
$$

Furthermore, using the fact that μ_t admits the density ρ_t, we can use $\frac{\partial}{\partial t}\rho_t = -\nabla \cdot (\rho_t v_t)$. For a more formal treatment in terms of measures, see e.g. Chapter 10 of Ambrosio et al. 2008.

This gives:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} KL(\mu_t||\pi) &= \int \log\left(\frac{\rho_t}{p}\right) \frac{\partial}{\partial t}\rho_t \mathrm{d}x \\
&= -\int \left(\log\left(\frac{\rho_t}{p}\right) \nabla \cdot (\rho_t v_t)\right) \mathrm{d}x \\
&= \int (\rho_t v_t) \cdot \nabla \log\left(\frac{\rho_t}{p}\right) \mathrm{d}x - \oint_{||x||\to\infty} \log\left(\frac{\rho_t}{p}\right) (\rho_t v_t) \cdot \boldsymbol{n} \mathrm{d}S(x) \\
&= \int (\rho_t v_t) \cdot \nabla \log\left(\frac{\rho_t}{p}\right) \mathrm{d}x \\
&= \langle v_t, \nabla \log\left(\frac{\rho_t}{p}\right) \rangle_{L^2(\mu_t)},
\end{aligned}$$

where in the third line we have used the divergence theorem (or an integration by parts). This gives the desired result. □

B Figures

B.1 Airline passenger data figure

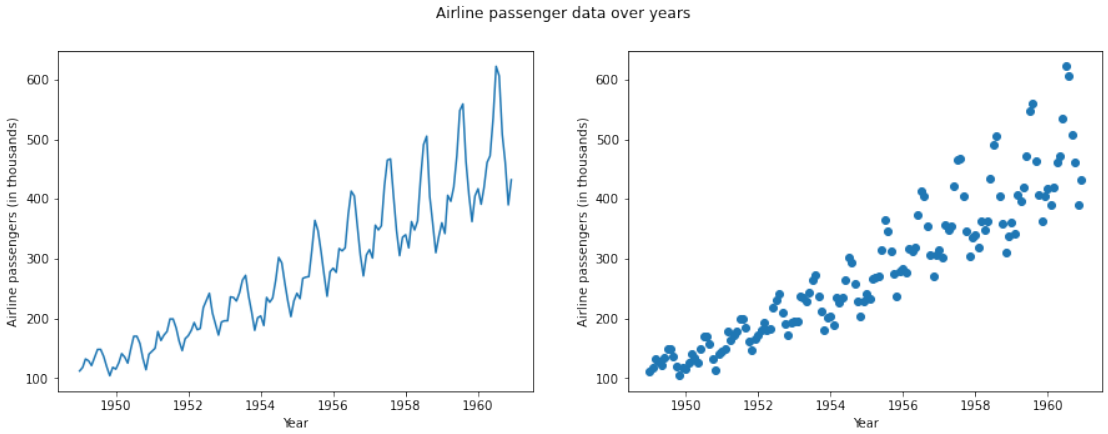

Figure 7: Data for the number of airline passengers, obtained every month in the displayed years.

C Calculations

C.1 Calculation of MAP estimation

We have that $\theta \sim \mathcal{N}(o, \frac{1}{\lambda} I)$ and that $Y|X \sim \mathcal{N}(f_\theta(x), \sigma_\theta^2 I)$. This gives the following formulas for their densities:

$$p(\theta) = \frac{1}{(2\pi)^{\frac{dim(\theta)}{2}} \left(\frac{1}{\lambda}\right)^{\frac{dim(\theta)}{2}}} \exp\left(-\frac{\lambda}{2}\theta^T\theta\right),$$
$$p(y|x,\theta) = \frac{1}{(2\pi)^{\frac{p_y}{2}} (\sigma_\theta^2(x))^{\frac{p_y}{2}}} \exp\left(-\frac{1}{2\sigma_\theta^2(x)}||y - f_\theta(x)||_2^2\right).$$

We also had $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where we assume the x_i are non-stochastic and the data is independent and identically distributed according to $Y|X \sim \mathcal{N}(f_\theta(x), \sigma_\theta^2 I)$. Hence, $p(\mathcal{D}|\theta) = p((x_1, y_1), \ldots, (x_n, y_n)|\theta) = \prod_{i=1}^n p(y_i|x_i, \theta)$. Let us now try to find the MAP estimate for θ:

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_\theta \{p(\theta|\mathcal{D})\} \\
&= \arg\max_\theta \{\log p(\theta|\mathcal{D})\} \\
&= \arg\max_\theta \{\log p(\theta) + \log p(\mathcal{D}|\theta) - \log p(\mathcal{D})\} \\
&= \arg\max_\theta \{\log p(\theta) + \log p(\mathcal{D}|\theta)\} \\
&= \arg\max_\theta \left\{-\frac{\lambda}{2}\theta^T\theta - \frac{dim(\theta)}{2}\log 2\pi - \frac{dim(\theta)}{2}\log\frac{1}{\lambda} + \log p(\mathcal{D}|\theta)\right\} \\
&= \arg\max_\theta \left\{-\frac{\lambda}{2}\theta^T\theta + \log p(\mathcal{D}|\theta)\right\}, \quad \text{as } dim(\theta) \text{ is fixed,} \\
&= \arg\max_\theta \left\{-\frac{\lambda}{2}\theta^T\theta - \sum_{i=1}^n \left(\frac{1}{2\sigma_\theta^2(x_i)}||y_i - f_\theta(x_i)||_2^2 + \frac{p_y}{2}\log\sigma_\theta^2(x_i) + \frac{p_y}{2}\log 2\pi\right)\right\} \\
&= \arg\max_\theta \left\{-\frac{\lambda}{2}||\theta||_2^2 - \frac{1}{2}\sum_{i=1}^n \left(\frac{1}{\sigma_\theta^2(x_i)}||y_i - f_\theta(x_i)||_2^2 + p_y \log\sigma_\theta^2(x_i)\right)\right\} \\
&= \arg\min_\theta \left\{\frac{\lambda}{2}||\theta||_2^2 + \frac{1}{2}\sum_{i=1}^n \left(\frac{1}{\sigma_\theta^2(x_i)}||y_i - f_\theta(x_i)||_2^2 + p_y \log\sigma_\theta^2(x_i)\right)\right\}.
\end{aligned}$$

The term in the minimization is exactly equal to the loss function in equation (8).

C.2 Showing the partial integration for the kernelized SVGD wasserstein gradient

We work out the partial integration from equation (72), which was not worked out in either Korba, Salim, et al. 2020 or Chewi et al. 2020. Denoting the density of μ_t as ρ_t and the density of π as p. Furthermore, we assume that $\lim_{||x||\to\infty} k(x, \cdot)\rho_t(x)$ is the zero function, denoted simply as 0. Remember that $\mathcal{K}_{\mu_t}\nabla_{W_2}\mathcal{F}(\mu_t) : \mathbb{R}^d \to \mathbb{R}^d$. We have the following chain of equalities:

$$
\begin{aligned}
\mathcal{K}_{\mu_t} \nabla_{W_2} \mathcal{F}(\mu_t) &= \int \nabla \log \left(\frac{\rho_t}{p} \right)(x) k(\cdot, x) \mathrm{d}\mu_t(x) \\
&= \int (\nabla \log(\rho_t(x))) \rho_t(x) k(\cdot, x) \mathrm{d}x - \int \nabla \log(p(x)) k(x, \cdot) \mathrm{d}\mu_t(x) \\
&= \int (\nabla \rho_t(x)) k(x, \cdot) \mathrm{d}x - \int \nabla \log(p(x)) k(x, \cdot) \mathrm{d}\mu_t(x) \\
&= \rho_t(x) k(x, \cdot) \Big|_{||x|| \to \infty} - \int \rho_t(x) \nabla_x k(x, \cdot) \mathrm{d}x - \int \nabla \log(p(x)) k(x, \cdot) \mathrm{d}\mu_t(x) \\
&= 0 - \int \nabla_x k(x, \cdot) \mathrm{d}\mu_t(x) - \int \nabla \log(p(x)) k(x, \cdot) \mathrm{d}\mu_t(x) \\
&= - \int \left(\nabla \log(p(x)) k(x, \cdot) + \nabla_x k(x, \cdot) \right) \mathrm{d}\mu_t(x).
\end{aligned}
$$

In the fourth line we have used an integration by parts. This gives the result.

D Complementary information

D.1 Kernelized Stein discrepancy

In this section we will state Definition 3.1 and show Proposition 3.3 from Liu, Lee, et al. 2016 to prove that KSD is such that under suitable positive definiteness conditions we have $S(q, p) \geq 0$ and $S(q, p) = 0$ if and only if $q = p$.

We will work with Definition A.1 and give another definition needed to state the result in Proposition D.2.

Definition D.1 (From Liu, Lee, et al. 2016)**.** A kernel $(x, x') \mapsto k(x, x')$ is integrally strictly positive definite, if for any function f which satisfies $0 \leq ||f||_2^2 < \infty$,

$$
\int_{\mathcal{X}} f(x) k(x, x') f(x') \mathrm{d}x \mathrm{d}x' > 0.
$$

Proposition D.2 (From Liu, Lee, et al. 2016)**.** *Define* $\boldsymbol{f}$ *as* $x \mapsto \boldsymbol{f}_{q,p}(x) = q(x)(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))$*. Assume the kernel function* k *is integrally strictly positive definite and* q, p *are continuous densities with* $||\boldsymbol{f}_{q,p}|| < \infty$*. Then we have that* $S(q, p) \geq 0$ *and* $S(q, p) = 0$ *if and only if* $q = p$*.*

Proof. Using the definition of KSD: $S(q, p) = \mathrm{E}_{x,x' \sim q}[\boldsymbol{\delta}_{p,q}(x)^T k(x, x') \boldsymbol{\delta}_{p,q}(x')]$, we have by the definition of the kernel function k being integrally strictly positive definite that for $q \neq p$:

$$
\begin{aligned}
S(q, p) &= \mathrm{E}_{x,x' \sim q}[\boldsymbol{\delta}_{p,q}(x)^T k(x, x') \boldsymbol{\delta}_{p,q}(x')] \\
&= \int_{\mathcal{X}} (q(x)(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)))^T k(x, x') q(x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x')) \mathrm{d}x \mathrm{d}x' \\
&> 0.
\end{aligned}
$$

If $q = p$, then $S(q, p) = 0$, as $\boldsymbol{\delta}_{p,q} = 0$. □

D.2 Kernelized Stein discrepancy: different definitions

Let us start with a specific definition of a Stein operator.

Definition D.3 (Definition from Gorham and Mackey 2017)**.** Consider some generic probability measure μ and a Stein operator $\mathcal{T}$ mapping functions $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$ from a domain $\mathcal{G}$ to real-valued functions $\mathcal{T}\boldsymbol{\phi}$ such that the following property holds:

$$\mathrm{E}_{X\sim\mu}[(\mathcal{T}\boldsymbol{\phi})(X)] = 0, \quad \text{for all } \boldsymbol{\phi} \in \mathcal{G}.$$

For any such Stein operator and Stein set $\mathcal{G}$, define the Stein discrepancy as:

$$\mathcal{S}(\mu, \mathcal{T}, \mathcal{G}) := \sup_{\boldsymbol{\phi}\in\mathcal{G}} |\,\mathrm{E}_\mu[(\mathcal{T}\boldsymbol{\phi})(X)]|.$$

Remark D.4. We can connect this definition with a previous result in Lemma 3.4, as the Stein operator $\mathcal{T}$ resembles the 'earlier' definition of the Stein operator $\mathcal{A}_p : \boldsymbol{\phi} \mapsto \boldsymbol{s}_p\boldsymbol{\phi}^T + \nabla\boldsymbol{\phi}$, with $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$. However, the Stein operator $\mathcal{T}$ is a scalar-valued function. Let us use $\mathcal{T}$ as follows, $\boldsymbol{\phi} \mapsto \mathcal{T}\boldsymbol{\phi} = \boldsymbol{s}_p^T\boldsymbol{\phi} + \nabla\cdot\boldsymbol{\phi}$. In this way it resembles $\mathcal{A}_p$, but now yielding a scalar-valued function. In this way we have made a Stein operator that maps vector-valued functions to real-valued functions $\mathcal{T}\boldsymbol{\phi}$. Furthermore, this Stein set $\mathcal{G}$ can be identified with the Stein class of the density of p.

Definition D.5 (Definition from Gorham and Mackey 2017)**.** The kernel Stein set $\mathcal{G}_{k,||\cdot||}$ is the set of vector-valued functions $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_d]^T$ with $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$ such that each component function ϕ_j belongs to the RKHS $\mathcal{H}$ with associated reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and induced norm $||\cdot||_{\mathcal{H}}$. The kernel Stein set is given as:

$$\mathcal{G}_{k,||\cdot||} = \{\boldsymbol{\phi} = [\phi_1, \cdots, \phi_d]^T \mid ||v||^* \leq 1 \text{ for } v_j = ||\phi_j||_{\mathcal{H}}\},$$

with $||\cdot||^*$ denoting the dual norm, associated to the norm $||\cdot||$. This dual norm is defined as $||a||^* := \sup_{b\in\mathbb{R}^d, ||b||=1} \langle a, b\rangle$ for vectors $a \in \mathbb{R}^d$.

Remark D.6. The dual with respect to the usual Euclidean norm on $\mathbb{R}^d$ is again the Euclidean norm. Let us show this, by considering the usual Euclidean inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$ and the corresponding norm $||\cdot||_2$. Take any $a, b \in \mathbb{R}^d$ with $||b||_2 = 1$. By the Cauchy-Schwarz inequality, $|\langle a, b\rangle| \leq ||a||_2||b||_2 = ||a||_2$, so $||a||_2^* \leq ||a||_2\ \forall a \in \mathbb{R}^d$. Let us now show that this upper bound is achieved. Take $b = a/||a||_2$ (and observe it has a norm of one), then $\langle a, b\rangle = \langle a, a/||a||_2\rangle = \frac{||a||_2^2}{||a||_2} = ||a||_2$. Hence, an upper bound is achieved and thus the supremum is equal to it. For the Euclidean norm we now have that the dual of this norm is equal to itself. Then, $||v||_2 \leq 1$ for $v_j = ||\phi_j||_{\mathcal{H}}$ is equal to $\sum_{j=1}^d ||\phi_j||_{\mathcal{H}}^2 = ||\boldsymbol{\phi}||_{\mathcal{H}^d}^2 \leq 1$. In turn, this shows that the set $\mathcal{G}_{k,||\cdot||_2}$ can be characterised as:

$$\mathcal{G}_{k,||\cdot||_2} = \{\boldsymbol{\phi} = [\phi_1, \cdots, \phi_d]^T \mid \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\}.$$

From now on we will work with the specific Stein operator $\boldsymbol{\phi} \mapsto \mathcal{T}_p\boldsymbol{\phi} = \boldsymbol{s}_p^T\boldsymbol{\phi} + \nabla\cdot\boldsymbol{\phi}$ for functions $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$. Together with the kernel Stein set $\mathcal{G}_{k,||\cdot||}$ this forms the kernel Stein discrepancy

(KSD), denoted $\mathcal{S}(\mu, \mathcal{T}_p, \mathcal{G}_{k,||\cdot||})$. If $||\cdot|| = ||\cdot||_2$, then it corresponds to the same definition of KSD as in Definition 3.6, except for a square root.

In fact, this definition of $\mathcal{S}(\mu, \mathcal{T}_p, \mathcal{G}_{k,||\cdot||_2})$ is the same as the definition of $D(\mu||\nu_p)$, where p is the density of some target measure ν_p. Here,

$$D(\mu||\nu_p) = \max\{\mathrm{E}_\mu[\mathrm{tr}(\mathcal{A}_p\boldsymbol{\phi})] \quad | \quad \boldsymbol{\phi} \in \mathcal{H}^d, ||\boldsymbol{\phi}||_{\mathcal{H}^d} \leq 1\} = \mathcal{S}(\mu, \mathcal{T}_p, \mathcal{G}_{k,||\cdot||_2}).$$

D.3 First variation

We defined

$$x \mapsto \frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x) := -\nabla \cdot F_p(x, \rho(x), \nabla\rho(x)) + F_z(x, \rho(x), \nabla\rho(x))$$

as the first variation of the functional $\mathcal{F}$, with the underlying assumption that our $\mathcal{F}$ can be written as follows:

$$\mathcal{F}(\rho) = \int_{\mathbb{R}^d} F(x, \rho(x), \nabla\rho(x))\mathrm{d}x,$$

with a smooth function $F = F(x, z, p) : \mathbb{R}^d \times [0, \infty) \times \mathbb{R}^d \to \mathbb{R}$. Let us motivate this definition of the first variation. Consider the following:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho_t) &= \int \left(F_z(x, \rho_t(x), \nabla\rho_t(x))\partial_t\rho_t(x) + F_p(x, \rho_t, \nabla\rho_t(x)) \cdot \nabla(\partial_t\rho_t)(x)\right)\mathrm{d}x \\
&= \int \left(F_z(x, \rho_t(x), \nabla\rho_t(x)) - \nabla \cdot F_p(x, \rho_t, \nabla\rho_t(x))\right)\partial_t\rho_t(x)\mathrm{d}x \\
&= \int \frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x)\partial_t\rho_t\mathrm{d}x.
\end{aligned}$$

So we can identify the integrand term between brackets as the first variation, where, loosely speaking, the first variation $\frac{\delta\mathcal{F}(\rho)}{\delta\rho}$ can be seen as representing the derivative of $\mathcal{F}$ in the following sense:

$$\mathcal{F}'(\rho)(h) = \int \frac{\delta\mathcal{F}(\rho)}{\delta\rho}(x)h(x)\mathrm{d}x,$$

for any (test) function h and $\mathcal{F}'(\rho)$ informally denoting a derivative of $\mathcal{F}$.