Stochastic Backpropagation through Fourier Transforms

Amine Echraibi, Joachim Flocon-Cholet, Stéphane Gosselin
Orange Labs
Lannion, France
{amine.echraibi, joachim.floconcholet, stephane.gosselin}@orange.com

Sandrine Vaton
IMT Atlantique
Brest, France
sandrine.vaton@imt-atlantique.fr

***Abstract*—Backpropagating gradients through random variables is at the heart of numerous machine learning applications. In this paper, we present a general framework for deriving stochastic backpropagation rules for any continuous distribution. Our approach exploits the link between the characteristic function and the Fourier transform, to transport the derivatives from the parameters of the distribution to the random variable. Our method generalizes previously known estimators, and results in new estimators for the gamma, beta, Dirichlet and Laplace distributions. Furthermore, we show that the classical deterministic backproapagation rule in neural networks, is a special case of stochastic backpropagation with Dirac distributions, thus providing a link between probabilistic graphical models and neural networks.**

***Index Terms*—Stochastic backpropagation, Variational inference.**

I. Introduction

Deep neural networks with stochastic hidden layers have become crucial in multiple domains, such as generative modeling [1]–[3], deep reinforcement learning [4], and attention mechanisms [5]. The difficulty encountered in training such models arises in the computation of gradients for functions of the form $\mathcal{L}(\theta) := \mathbb{E}_{\mathbf{z}\sim p_\theta}[f(\mathbf{z})]$ with respect to the parameters θ, thus needing to backpropagate the gradient through the random variable $\mathbf{z}$ [6]. One of the first and most used methods is the score function or reinforce method [7], [8], that requires the computation and estimation of the derivative of the log probability function. For high dimensional applications however, it has been noted that reinforce gradients have high variance, making the training process unstable [2].

Recently, significant progress has been made in tackling the variance problem. The first class of approaches dealing with continuous random variables are reparameterization tricks. In that case a standardization function is introduced, that separates the stochasticity from the dependency on the parameters θ. This enables to transport the derivative inside the expectation and to sample from a fixed distribution, resulting in low variance gradient [1], [2], [9]–[12]. The second class of approaches concerns discrete random variables, for which a direct reparameterization is not known. The first solution uses the score function gradient with control variate methods to reduce its variance [3], [13]. The second consists in introducing a continuous relaxation admitting a reparameterization trick of the discrete random variable, thus being able to backpropagate low-variance reparameterized gradients by sampling from the concrete distribution [14]–[17].

Although recent developments have advanced the state-of-the-art in terms of variance reduction and performance, stochastic backpropagation (i.e computing gradients through random variables) still lacks theoretical foundation. In particular, the following questions remain open: How to develop stochastic backpropagation rules, where the derivative is transferred explicitly to the function f for a broader range of distributions? And can the deterministic case be interpreted in the sense of stochastic backpropagation? In this paper, we provide a new method to address these questions, and our main contributions are the following:

- We present a theoretical framework based on the link between the multivariate Fourier transform and the characteristic function, that provides a standard method for deriving stochastic backpropagation rules, for **any** continuous distribution.
- We show that deterministic backpropagation can be interpreted as a special case of stochastic backpropagation, where the probability distribution p_θ is a Dirac delta distribution.
- We generalize previously known estimators, and provide new stochastic backpropagation rules for the special cases of the Laplace, gamma, beta, and Dirichlet distributions.
- We demonstrate experimentally that the resulting new estimators are competitive with state-of-the art methods on simple tasks.

II. Background & Preliminaries

Let (E, λ) be a d-dimensional measure space equipped with the standard inner product, and f be a square summable positive real valued function on E, that is, $f\colon E \to \mathbb{R}_+$, with $\int_E |f(z)|^2 \lambda(dz) < \infty$. Let p_θ be an arbitrary parameterized probability density on the space E. We denote by φ_θ its characteristic function, defined as: $\varphi_\theta(\omega) := \mathbb{E}_{\mathbf{z}\sim p_\theta}[e^{i\omega^T \mathbf{z}}]$. We denote by $\hat{f}$ the Fourier transform of the function f defined as:

$$\hat{f}(\omega) := \mathcal{F}\{f\}(\omega) = \int_E f(z) e^{-i\omega^T z} \lambda(dz). \tag{1}$$

The inverse Fourier transform is given in this case by:

$$f(z) := \mathcal{F}^{-1}\{\hat{f}\}(z) = \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^T z} \mu(d\omega), \tag{2}$$

where $\mu(d\omega)$ represents the measure in the Fourier domain. In this paper we treat the case where E is a continuous sub domain of $\mathbb{R}^d$, thus, $\mu(d\omega) = \frac{d\omega}{(2\pi)^d}$. Throughout the paper, we reserve the letter i to denote the imaginary unit: $i^2 = -1$. To denote higher order derivatives of the function f, we use the multi-index notation [18]. For a multi-index $n = (n_1, ..., n_d) \in \mathbb{N}^d$, we define:

$$\partial_z^n := \frac{\partial^{|n|}}{\partial z_1^{n_1} ... \partial z_d^{n_d}} \quad \text{where} \quad |n| = \sum_{j=1}^{d} n_j, \quad \omega^n := \prod_{j=1}^{d} \omega_j^{n_j}.$$

To clarify the multi-index notation, let us consider the example where $d = 3$, and $n = (1, 0, 2)$, in this case:

$$\partial_z^n = \frac{\partial^3}{\partial z_1 \partial z_3^2} \quad \text{and,} \quad \omega^n = \omega_1 \omega_3^2.$$

The objective is to derive stochastic backpropagation rules, similar to that of [2], for functions of the form: $\mathcal{L}(\theta) := \mathbb{E}_{\mathbf{z} \sim p_\theta}[f(\mathbf{z})]$, for any continuous distribution p_θ.

III. Fourier Stochastic Backpropagation

Stochastic backpropagation rules similar to that of [2] can in fact be derived for any continuous distribution, under certain conditions on the characteristic function. In the following theorem we present the main result of our paper concerning the derivation of Fourier stochastic backpropagation rules.

Theorem 1. *Let $f \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R}_+)$, under the condition that $\nabla_\theta \log \varphi_\theta$ is a holomorphic function of $i\omega$, then there exists a unique set of real numbers $\{a_n(\theta)\}_{n \in \mathbb{N}^d}$ such that:*

$$\nabla_\theta \mathcal{L} = \sum_{|n| \geq 0} a_n(\theta) \mathbb{E}_{\mathbf{z} \sim p_\theta} \left[\partial_z^n f(\mathbf{z}) \right]. \tag{3}$$

Where $\{a_n(\theta)\}_{n \in \mathbb{N}^d}$ are the Taylor expansion coefficients of $\nabla_\theta \log \varphi_\theta(\omega)$:

$$\nabla_\theta \log \varphi_\theta(\omega) = \sum_{|n| \geq 0} a_n(\theta)(i\omega)^n. \tag{4}$$

Proof. Let us rewrite $\mathcal{L}$ in terms of $\hat{f}$:

$$\begin{aligned} \mathcal{L}(\theta) &= \int p_\theta(z) f(z) \lambda(dz) \\ &= \int p_\theta(z) \mathcal{F}^{-1}[\hat{f}](z) \lambda(dz) \\ &= \int_{\mathbb{R}^d} \hat{f}(\omega) \int_E p_\theta(z) e^{i\omega^T z} \lambda(dz) \mu(d\omega) \\ &= \int_{\mathbb{R}^d} \hat{f}(\omega) \varphi_\theta(\omega) \mu(d\omega). \end{aligned} \tag{5}$$

By introducing the derivative under the integral sign, and using the reinforce trick [8] applied to φ_θ, where $\nabla_\theta \varphi_\theta(\omega) = \varphi_\theta(\omega) \nabla_\theta \log \varphi_\theta(\omega)$, (5) becomes:

$$\nabla_\theta \mathcal{L} = \int_{\mathbb{R}^d} \hat{f}(\omega) \varphi_\theta(\omega) \nabla_\theta \log \varphi_\theta(\omega) \mu(d\omega). \tag{6}$$

Under analyticity conditions of the gradient of the log characteristic function, we can expand the gradient term $\nabla_\theta \log \varphi_\theta(\omega)$, in terms of Taylor series around zero as:

$$\nabla_\theta \log \varphi_\theta(\omega) = \sum_{|n| \geq 0} a_n(\theta)(i\omega)^n. \tag{7}$$

Putting everything together, and replacing the characteristic function by its expression, the gradient of $\mathcal{L}$ becomes:

$$\begin{aligned} \nabla_\theta \mathcal{L} = \int_{\mathbb{R}^d} \hat{f}(\omega) \int_E p_\theta(z) e^{i\omega^T z} \\ \times \sum_{|n| \geq 0} a_n(\theta)(i\omega)^n \mu(d\omega) \lambda(dz). \end{aligned} \tag{8}$$

By rearranging the sums using Fubini's theorem a second time, we obtain the following expression for the gradient:

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \mathbb{E}_{\mathbf{z} \sim p_\theta} \left[\mathcal{F}^{-1} \left\{ \omega \mapsto \sum_{|n| \geq 0} a_n(\theta)(i\omega)^n \hat{f}(\omega) \right\} (\mathbf{z}) \right] \\ &= \sum_{|n| \geq 0} a_n(\theta) \mathbb{E}_{\mathbf{z} \sim p_\theta} \left[\mathcal{F}^{-1} \left\{ \omega \mapsto (i\omega)^n \hat{f}(\omega) \right\} (\mathbf{z}) \right] \\ &= \sum_{|n| \geq 0} a_n(\theta) \mathbb{E}_{\mathbf{z} \sim p_\theta} \left[\partial_z^n f(\mathbf{z}) \right]. \end{aligned} \tag{9}$$

□

IV. Applications of Fourier Stochastic Backpropagation

Following from the previous section, we derive the stochastic backpropagation estimators for certain commonly used distributions.

The multivariate Gaussian distribution: In this case $p_\theta(z) = \mathcal{N}(z; \mu_\theta, \Sigma_\theta)$. The log characteristic function is given by: $\log \varphi_\theta(\omega) = i\mu_\theta^T \omega + \frac{1}{2} \text{Tr}\left[\Sigma_\theta i^2 \omega \omega^T\right]$. Thus by applying theorem 1, we recover the stochastic backpropagation rule of [2]:

$$\begin{aligned} \nabla_\theta \mathcal{L} = \left(\frac{\partial \mu_\theta}{\partial \theta} \right)^T \mathbb{E}_{\mathbf{z} \sim p_\theta} \left\{ \nabla_z f(\mathbf{z}) \right\} \\ + \frac{1}{2} \text{Tr} \left[\left(\frac{\partial \Sigma_\theta}{\partial \theta} \right) \mathbb{E}_{\mathbf{z} \sim p_\theta} \left\{ \nabla_z^2 f(\mathbf{z}) \right\} \right], \end{aligned} \tag{10}$$

where, ∇_z and ∇_z^2, represent the gradient and hessian operators.

The multivariate Dirac distribution: $p_\theta(z) = \delta_{a_\theta}(z)$, the log characteristic function of the Dirac distribution is given by: $\log \varphi_\theta(\omega) = i\omega^T a_\theta$. Thus the stochastic backpropagation rule of the Dirac is given by:

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \left(\frac{\partial a_\theta}{\partial \theta} \right)^T \mathbb{E}_{\mathbf{z} \sim \delta_{a_\theta}} \left[\nabla_z f(\mathbf{z}) \right] \\ &= \left(\frac{\partial a_\theta}{\partial \theta} \right)^T \nabla_z f(a_\theta), \end{aligned} \tag{11}$$

resulting in the classical backpropagation rule. In other words, the deterministic backpropagation rule is a special case of stochastic backpropagation where the distribution is a

Dirac delta distribution. This result provides a link between probabilistic graphical models and classical neural networks. Namely, when using neural networks we are indirectly using a probabilistic graphical model and making the strong assumption that the hidden layers follow a parameterized Dirac distribution knowing the previous layer.

The exponential distribution: $p_\theta(z) = \mathcal{E}(z;\lambda_\theta)$, in this case the log characteristic function is given by: $\log\varphi_\theta(\omega) = -\log\left(1-\frac{i\omega}{\lambda_\theta}\right)$, using the Taylor series expansion for the logarithm, we get the following stochastic backpropagation rule for the exponential distribution:

$$\nabla_\theta\mathcal{L} = -\frac{1}{\lambda_\theta}\frac{\partial\lambda_\theta}{\partial\theta}\sum_{n=1}^{\infty}\frac{n}{\lambda_\theta^n}\mathbb{E}_{\mathbf{z}\sim p_\theta}\left[\frac{d^n f}{dz^n}(\mathbf{z})\right] \tag{12}$$

The Laplace distribution: $p_\theta(z) = L(z;\mu_\theta,b_\theta)$, in this case the log characteristic function is the following: $\log\varphi_\theta(\omega) = i\mu_\theta\omega - \log(1+b_\theta^2\omega^2)$, using the Taylor series expansion for the function $x\mapsto\frac{1}{1-x}$, we get the following stochastic backpropagation rule for the Laplace distribution:

$$\begin{aligned}\nabla_\theta\mathcal{L} = &\frac{\partial\mu_\theta}{\partial\theta}\mathbb{E}_{\mathbf{z}}\left[\frac{df}{dz}(\mathbf{z})\right]\\ &+\frac{1}{b_\theta^2}\frac{\partial b_\theta^2}{\partial\theta}\sum_{n=1}^{\infty}b_\theta^{2n}\mathbb{E}_{\mathbf{z}}\left[\frac{d^{2n}f}{dz^{2n}}(\mathbf{z})\right].\end{aligned} \tag{13}$$

The gamma distribution: $p_\theta(z) = \Gamma(z;k_\theta,\mu_\theta)$, the log characteristic function of the Gamma distribution is given by: $\log\varphi_\theta(\omega) = -k_\theta\log(1-i\mu_\theta\omega)$. By expanding it using Taylor series of the logarithm function, we obtain the following stochastic backpropagation rule:

$$\nabla_\theta\mathcal{L} = \sum_{n=1}^{\infty}\left[\frac{1}{n}\frac{\partial k_\theta}{\partial\theta}+\frac{k_\theta}{\mu_\theta}\frac{\partial\mu_\theta}{\partial\theta}\right]\mu_\theta^n\mathbb{E}_{\mathbf{z}\sim p_\theta}\left[\frac{d^n f}{dz^n}(\mathbf{z})\right]. \tag{14}$$

The estimator of (14) gives a stochastic backpropagation rule for the gamma distribution and, hence also applies by extension to the special cases of the Erlang, and chi-squared distributions.

The beta distribution: $p_\theta(z) = \text{Beta}(z;\alpha_\theta,\beta_\theta)$, in this case the characteristic function is the confluent hypergeometric function: $\varphi_\theta(\omega) = {}_1F_1(\alpha_\theta;\alpha_\theta+\beta_\theta;i\omega)$. A series expansion of the gradient of the log of this function is not trivial to derive. However, we can use the parameterization linking the gamma and beta distributions to derive a stochastic backpropagation rule. Indeed, if $\boldsymbol{\zeta}_1\sim\Gamma(\alpha_\theta,1)$ and $\boldsymbol{\zeta}_2\sim\Gamma(\beta_\theta,1)$, then $\mathbf{z} = g(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2) = \frac{\boldsymbol{\zeta}_1}{\boldsymbol{\zeta}_1+\boldsymbol{\zeta}_2}\sim\text{Beta}(\alpha_\theta,\beta_\theta)$. By substituting in the gamma stochastic backpropagation rule, we obtain:

$$\begin{aligned}\nabla_\theta\mathcal{L} = \sum_{n=1}^{\infty}\frac{1}{n}\Bigg\{&\frac{\partial\alpha_\theta}{\partial\theta}\mathbb{E}_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}\left[\frac{\partial^n f}{\partial\zeta_1^n}\left(\frac{\boldsymbol{\zeta}_1}{\boldsymbol{\zeta}_1+\boldsymbol{\zeta}_2}\right)\right]\\ &+\frac{\partial\beta_\theta}{\partial\theta}\mathbb{E}_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}\left[\frac{\partial^n f}{\partial\zeta_2^n}\left(\frac{\boldsymbol{\zeta}_1}{\boldsymbol{\zeta}_1+\boldsymbol{\zeta}_2}\right)\right]\Bigg\}.\end{aligned} \tag{15}$$

The Dirichlet distribution: $p_\theta(z) = \text{Dir}(z;K,\alpha_\theta)$, following the same procedure, as for the beta distribution and using the following parameterization: $\mathbf{z}_k = \frac{\boldsymbol{\zeta}_k}{\sum_{j=1}^K\boldsymbol{\zeta}_j}$ with, $\boldsymbol{\zeta}_k\sim\Gamma(\alpha_\theta^{(k)},1)$, we obtain:

$$\begin{aligned}\nabla_\theta\mathcal{L} = \sum_{n=1}^{\infty}\frac{1}{n}\Bigg\{&\sum_{k=1}^{K}\frac{\partial\alpha_\theta^{(k)}}{\partial\theta}\\ &\times\mathbb{E}_{\boldsymbol{\zeta}_j\forall j}\left[\frac{\partial^n f}{\partial\zeta_k^n}\left(\frac{\boldsymbol{\zeta}_1}{\sum_{j=1}^K\boldsymbol{\zeta}_j},...,\frac{\boldsymbol{\zeta}_K}{\sum_{j=1}^K\boldsymbol{\zeta}_j}\right)\right]\Bigg\}.\end{aligned} \tag{16}$$

V. Tractable cases & Approximations of Fourier Stochastic Backpropagation

The Fourier stochastic backpropagation gradient as presented in previous sections presents two major computational bottlenecks for non-trivial distributions. The first is the computation of infinite series, and the second is evaluating higher order derivatives of the function f. Depending on the application, the function f could be chosen in order to bypass the computational bottlenecks. A trivial example, is if the higher order derivatives of the function f vanish at a certain order: $\partial_z^n f = 0$. Another example, is the exponential function $f(z) = \exp(\epsilon^T z)$. From the fact that it obeys the following partial differential equation $\frac{\partial f}{\partial z_j}(z) = \epsilon_j f(z)$, one can deduce that the stochastic backpropagation rule reduces in this case to:

$$\nabla_\theta\mathcal{L} = \nabla_\theta\log\varphi_\theta\left(\frac{\epsilon}{i}\right)\mathbb{E}_{\mathbf{z}\sim p_\theta}\left[f(\mathbf{z})\right] \tag{17}$$

In most real world applications however, the infinite sum will not often reduce to a tractable expression such as that of the exponential. An example of this case is the evidence lower bound of a generative model with Bernoulli observations. In this case, the natural solution is to truncate the sum up to a finite order. The assumption is that the components associated to higher frequencies of the spectrum of the gradient of the log characteristic function, do not contribute as much. And by analogy to the signal processing field, we apply a Low-pass filter to eliminate them. In this case the gradient of the log characteristic function of (7) becomes:

$$\nabla_\theta\log\varphi_\theta(\omega) = \sum_{n\leq N}a_n(\theta)(i\omega)^n + o((i\omega)^N). \tag{18}$$

VI. Experiments

In our experimental evaluations, we test the stochastic backpropagation estimators of equations (13) and (14) for the gamma and Laplace distributions. In the case of the gamma estimator, we use toy examples where we can derive exact stochastic backpropagation rules without truncating the infinite sum. As for the Laplace stochastic backpropagation rule, we test the estimator in the case of Bayesian logistic regression with Laplacian priors and variational posteriors on the weights. We compare our estimators with the pathwise [19], [20], and score function estimators, in addition to the weak reparameterization estimator in the gamma case [10]. We do not use control variates in our setup, the goal is to verify the exactness of the proposed infinite series estimators and how they compare to current state-of-the-art methods in simple settings. In all our experiments, we use the Adam optimizer to update the weights

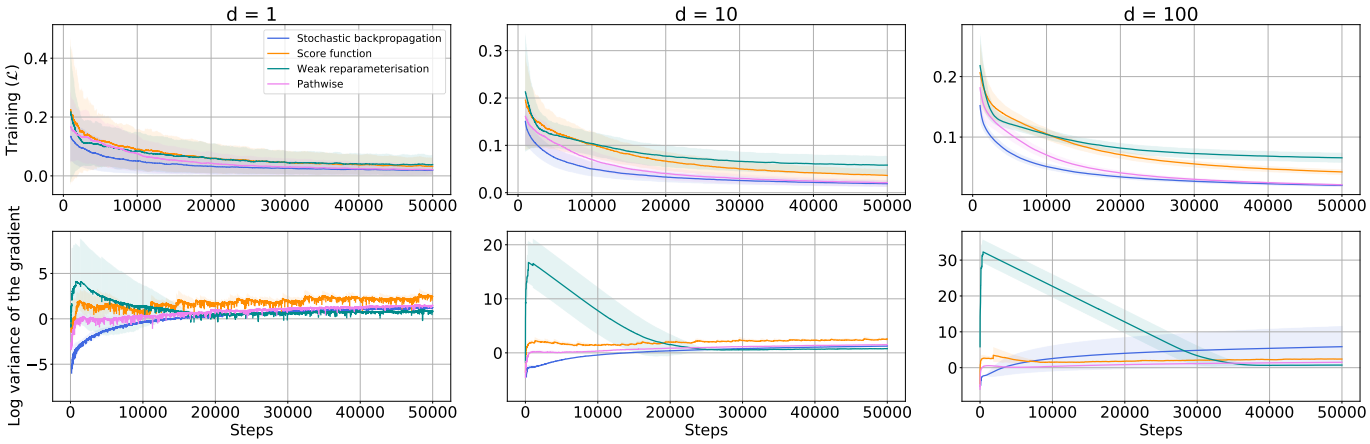

Fig. 1: Training loss and log variance of the gradients for the different estimators for $f(z) = \sum_{j=1}^{d}(z_j - \epsilon)^2$ for $d \in \{1, 10, 100\}$.

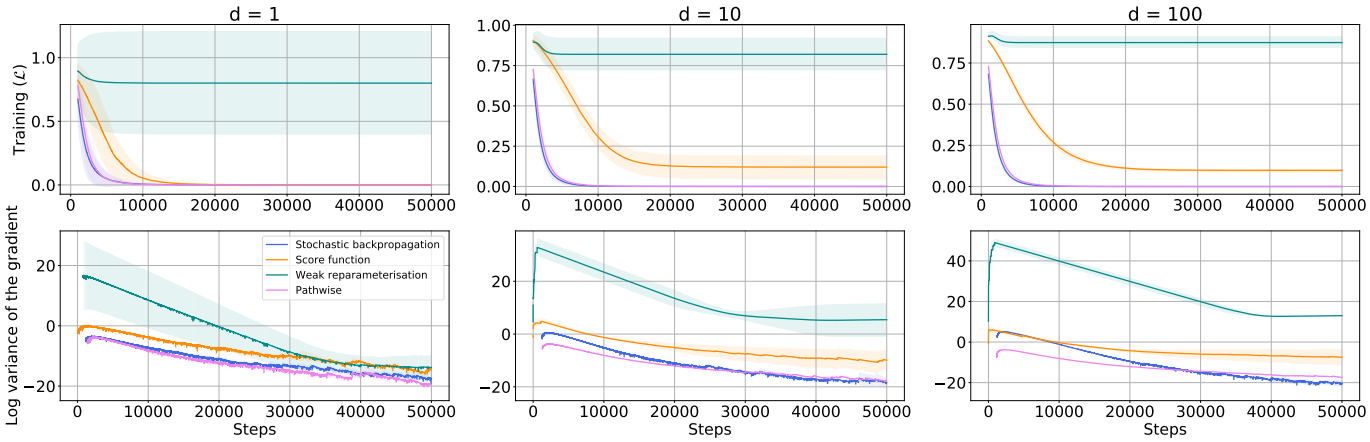

Fig. 2: Training loss and log variance of the gradients for the different estimators for $f(z) = \sum_{j=1}^{d} \exp(-\epsilon z_j)$ for $d \in \{1, 10, 100\}$.

[21], with a standard learning rate of 10^{-3}. In all the curves, we report the mean and standard deviation for all the metrics considered over 5 iterations.

A. Toy problems

In the toy problem setting, we test the gamma stochastic backpropagation rule following the same procedure as [22]. we consider the following cases:

Toy problem 1: $\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}\sim p_\theta}\left[||\mathbf{z} - \epsilon||^2\right]$, where $p_\theta(z) = \prod_{j=1}^{d} \Gamma(z_j; k_j, \mu_j)$, $\theta = \{k, \mu\}$, and $\epsilon = .49$. In this case, we only need to compute the first and second order derivatives of the function f.

Toy problem 2: $\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}\sim p_\theta}\left[\sum_{j=1}^{d} \exp(-\epsilon \mathbf{z}_j)\right]$, in this case, the infinite sum transfers to ϵ, which results in the following estimator: $\nabla_\theta \mathcal{L} = \nabla_\theta \log \varphi_\theta\,(i\epsilon)\, \mathbb{E}_{\mathbf{z}\sim p_\theta}\left[f(\mathbf{z})\right]$.

In figures 1 and 2 we report the training loss and log variance of the gradient across iterations of gradient descent for different values of the dimension $d \in \{1, 10, 100\}$. The stochastic backpropagation estimator converges to the minimal value in all cases faster than the other estimators and the variance of the gradient is competitive with the pathwise gradient.

B. Bayesian logistic regression with Laplacian Priors

We evaluate the Laplace stochastic backpropagation estimator using a Bayesian logistic regression model [23], similarly to [22]. In our case, we substitute the normal prior and posterior on the weights with Laplace priors and posteriors. We adopt the same notations as in [24], where the data, target and weight variables are respectively: $x_n \in \mathbb{R}^d$, $y_n \in \{-1, 1\}$, and $\mathbf{w}$. The probabilistic model in our case is the following:

$$p(w) = \prod_{j=1}^{d} L(w_j, 0, 1) \qquad p(y|\mathbf{x}, \mathbf{w}) = \sigma(y\mathbf{x}^T\mathbf{w}), \tag{19}$$

where σ represents the sigmoid function. We consider Laplacian variational posteriors of the form:

$$p_\theta(w) = \prod_{j=1}^{d} L(w_j, \mu_j, b_j), \tag{20}$$

with $\theta = \{\mu, b\}$. The evidence lower bound of a single sample is given by:

$$\mathcal{L}(x_n, y_n; \theta) = \mathbb{E}_{\mathbf{w}\sim p_\theta}\left[\log \sigma(y_n x_n^T \mathbf{w})\right] - \mathbb{D}_{KL}[p_\theta||p], \tag{21}$$

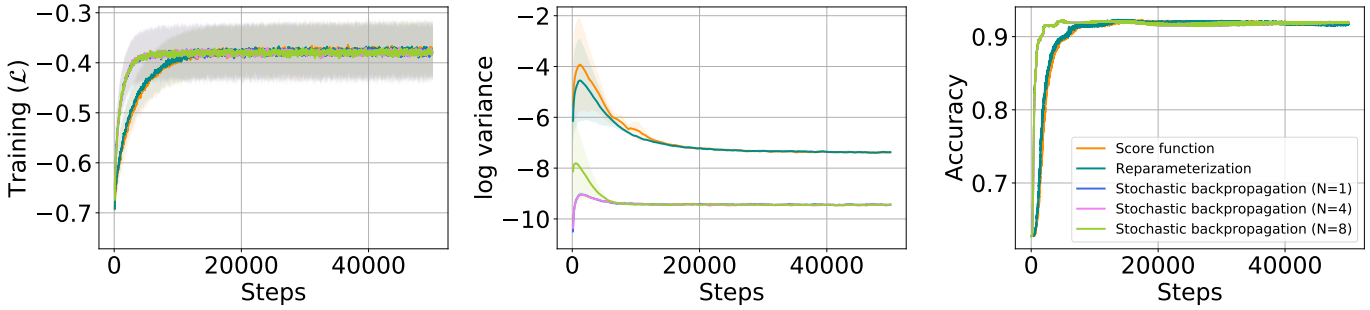

Fig. 3: Bayesian Logistic Regression with Laplacian priors

where the Kullback-Leibler divergence between the two Laplace distributions is the following:

$$\mathbb{D}_{KL}[p_\theta||p] = \sum_{j=1}^{d} \left\{ |\mu_j| + b_j e^{-\frac{|\mu_j|}{b_j}} - \log b_j - 1 \right\}. \quad (22)$$

We test the model on the UCI women's breast cancer dataset [25], with a batch size of 64 and 50 samples from the posterior to evaluate the expectation. In the case of the stochastic backpropagation estimator we truncate the infinite series for the scale parameter b of equation (13) to $N = 4$ and $N = 8$. In figure 3, we report the training evidence lower bound, the log variance of the gradient, and the accuracy computed on the entire dataset for the different estimators. The stochastic backpropagation estimator converges faster than the considered estimators and the variance is significantly lower. We also notice that the truncation level of the infinite series for the scale parameter has little effect on the outcome.

VII. Conclusion

In conclusion, in this paper we presented a new method to compute gradients through random variables for any probability distribution, by explicitly transferring the derivative to the random variable using the Fourier transform. Our approach, gives a framework to be applied for any distribution, where the gradient of the log characteristic function is analytic, resulting in a new broad family of stochastic backpropagation rules, that are unique for each distribution.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, 2014, pp. 1278–1286.

[3] A. Mnih and K. Gregor, "Neural variational inference and learning in belief networks," in *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, 2014, pp. II–1791.

[4] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[5] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[6] J. Schulman, N. Heess, T. Weber, and P. Abbeel, "Gradient estimation using stochastic computation graphs," in *Advances in Neural Information Processing Systems*, 2015, pp. 3528–3536.

[7] P. W. Glynn, "Optimization of stochastic systems via simulation," in *Proceedings of the 21st conference on Winter simulation*, 1989, pp. 90–105.

[8] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[9] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational bayes for non-conjugate inference," in *International conference on machine learning*, 2014, pp. 1971–1979.

[10] F. R. Ruiz, M. T. R. AUEB, and D. Blei, "The generalized reparameterization gradient," in *Advances in neural information processing systems*, 2016, pp. 460–468.

[11] C. Naesseth, F. Ruiz, S. Linderman, and D. Blei, "Reparameterization gradients through acceptance-rejection sampling algorithms," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 2017.

[12] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit reparameterization gradients," in *Advances in Neural Information Processing Systems*, 2018, pp. 441–452.

[13] S. Gu, S. Levine, I. Sutskever, and A. Mnih, "Muprop: Unbiased backpropagation for stochastic neural networks," 2016.

[14] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[15] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," 2016.

[16] G. Tucker, A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein, "Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models," in *Advances in Neural Information Processing Systems*, 2017, pp. 2627–2636.

[17] W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud, "Backpropagation through the void: Optimizing control variates for black-box gradient estimation," 2018.

[18] X. Saint Raymond, *Elementary introduction to the theory of pseudodifferential operators*. Routledge, 2018, ch. 1, pp. 2–3.

[19] M. Jankowiak and T. Karaletsos, "Pathwise derivatives for multivariate distributions," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 333–342.

[20] M. Jankowiak and F. Obermeyer, "Pathwise derivatives beyond the reparameterization trick," in *International Conference on Machine Learning*, 2018, pp. 2235–2244.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte carlo gradient estimation in machine learning," *arXiv preprint arXiv:1906.10652*, 2019.

[23] T. Jaakkola and M. Jordan, "A variational approach to bayesian logistic regression models and their extensions," in *Sixth International Workshop on Artificial Intelligence and Statistics*, vol. 82, no. 4, 1997.

[24] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[25] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml