Reparametrisation Gradient and Convergent SGD for Non-Differentiable Models via Smoothing: A Programming Language Approach

Anonymous Authors[1]

Abstract

It is well-known that the reparametrisation gradient estimator for non-differentiable models is biased. This may compromise correctness of gradient-based optimisation methods such as stochastic gradient descent (SGD) even when they converge. To formalise the problem, we consider a simple (higher-order, probabilistic) programming language with conditionals and a type system enforcing a mild restriction on the dependence on parameters. We endow our language with a *smoothed* (value) semantics parametrised by an accuracy coefficient, which yields an unbiased estimator. Finally, we present a novel variant of SGD, *Diagonalisation Stochastic Gradient Descent*, which enhances the accuracy whilst taking gradient steps, and we prove convergence to stationary points. Our estimator enjoys a similarly low variance as the reparametrisation gradient, while remaining unbiased for non-differentiable models. Experiments with our prototype implementation confirm the benefits of reduced variance and unbiasedness.

1. Introduction

In probabilistic programming and Bayesian machine learning we are often interested in minimising (or maximising) expectations of the form

$$\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[f(\boldsymbol{\theta}, \mathbf{z})] \tag{1}$$

i.e. to find parameters θ minimising (or maximising) that expectation. We are especially interested in scenarios in which f is directly expressed in a programming language. (Our ultimate goal is to target languages for constructing applications in numerical analysis and computational science such as Julia (Bezanson et al., 2017) or in Bayesian statistics and probabilistic programming such as Stan (Carpenter et al., 2017) and Pyro (Bingham et al., 2019).) Owing to the presence of conditionals, f may not be continuous, let alone differentiable. Also note that the distribution $q(\mathbf{z})$ (w.r.t. which the expectation is taken) is independent of the parameters θ.

An important example is the *evidence lower bound* (ELBO), which arises in variational inference for a reparameterisable variational family. Then f is the so-called *instantaneous ELBO*, $\log p(\mathbf{x}, \phi_{\boldsymbol{\theta}}(\mathbf{z})) - \log q_{\boldsymbol{\theta}}(\phi_{\boldsymbol{\theta}}(\mathbf{z}))$, and p is called the *model* and $q_{\boldsymbol{\theta}}$ the reparametrised *guide*, a member of the variational family. We reparametrise the latent variables $\boldsymbol{\alpha}$ in terms of a known base distribution (entropy source) via a diffeomorphic transformation $\phi_{\boldsymbol{\theta}}$ (such as a location-scale transformation or cumulative distribution function). E.g. if $q_{\boldsymbol{\theta}}(\alpha)$ is a Gaussian $\mathcal{N}(\alpha \mid \mu, \sigma^2)$ with $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ then the location-scale transformation using the standard normal as the base gives rise to the reparametrisation

$$\alpha \sim \mathcal{N}(\alpha \mid \mu, \sigma^2) \iff \alpha = \mu + \sigma\, z, \quad z \sim \mathcal{N}(0, 1)$$

The idea (often called "reparametrisation trick") is that we can now differentiate (by backpropagation) w.r.t. the parameters $\boldsymbol{\theta}$ of the variational distributions using a Monte Carlo simulation with draws from the base distribution. Thus, succinctly

$$\nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{\alpha\sim q_{\boldsymbol{\theta}}(\alpha)}[f(\boldsymbol{\theta}, \alpha)] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mu + \sigma z)]$$

(See (Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014; Rezende et al., 2014).)

Whilst it is less widely applicable than the SCORE estimator (Ranganath et al., 2014), the main benefit of the reparametrisation gradient estimator is a typically significantly lower variance, which is favourable for fast convergence.

Gradient Based Optimisation

Unfortunately, since $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[f(\boldsymbol{\theta}, \mathbf{z})]$ may not be convex, we cannot hope to always find global optima. We seek instead *stationary points*, where the gradient w.r.t. the parameters $\boldsymbol{\theta}$ vanishes.

In practice, variants of ***Stochastic Gradient Descent (SGD)*** are frequently employed, which converge to critical points

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

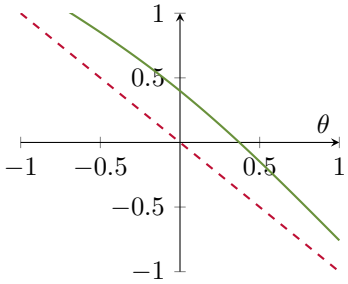


Figure 1. Dashed red: biased estimator $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\nabla_\theta f(\theta, z)\right]$, solid green: true gradient $\nabla_\theta \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[f(\theta, z)\right]$ for Example 1.1.

under certain conditions (Robbins and Monro, 1951). In its simplest form, SGD follows Monte Carlo estimates of the gradient in each step:

$$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \gamma_k \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta f\left(\boldsymbol{\theta}_k, \mathbf{z}_k^{(i)}\right)}_{\text{reparametrisation gradient estimator}}$$

where $\mathbf{z}_k^{(i)} \sim q\left(\mathbf{z}_k^{(i)}\right)$ and γ_k is the step size.

A necessary condition for convergence to critical points is unbiasedness of the gradient estimator. Unfortunately, it is well-known that the reparametrisation gradient estimator is biased for non-differentiable models (Lee et al., 2018), which are readily expressible in programming languages with conditionals:

Example 1.1. The counterexample in (Lee et al., 2018, Proposition 2), where the objective function is the ELBO for a non-differentiable model, can be simplified to

$$f(\theta, z) = -0.5 \cdot \theta^2 + \begin{cases} 0 & \text{if } z + \theta < 0 \\ 1 & \text{otherwise} \end{cases}$$

Observe that (see Figure 1):

$$\begin{aligned}\nabla_\theta \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[f(\theta, z)\right] &= -\theta + \mathcal{N}(-\theta \mid 0, 1) \\ &\neq -\theta = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\nabla_\theta f(\theta, z)\right]\end{aligned}$$

Unfortunately *this may compromise convergence to critical points or minimisers*: notably even if we can find a point where the gradient estimator vanishes, it may not be a critical point of the *original* problem (1).

Contributions

Our main contribution is the *provable* convergence of a novel variant of SGD, *Diagonalisation Stochastic Gradient Descent*, to stationary points for models defined by typable programs. The method takes gradient steps of a *smoothed* model whilst simulatiously enhancing the accuracy of the approximation in each iteration.

To formalise the approach, we present a simple (higher-order) programming language with conditionals, which is endowed with a smoothed (value) semantics. In particular, for *accuracy coefficient* $k \in \mathbb{N}$, we interpret conditionals using sigmoid functions as e.g.

$$[\![\mathbf{if}\ z + \theta < 0\ \mathbf{then}\ \underline{0}\ \mathbf{else}\ \underline{1}]\!]_k(\theta, z) := \sigma_k(z + \theta)$$

where $\sigma_k(x) := \sigma(\sqrt{k} \cdot x) = \frac{1}{1+\exp(-\sqrt{k}\cdot x)}$. We also devise a type system enforcing a mild restriction on the dependence of guards on parameters.

For the smoothed problems we obtain unbiased gradient estimators, which converge uniformly to the true gradient as the accuracy is improved.

Empirical studies show that our unbiased estimator perform comparably to the unbiased correction of the reparametrised gradient estimator (Lee et al., 2018), exhibiting a similar convergence. However our estimator is simpler, faster, and attains orders of magnitude (2 to 3,000 x) reduction in work-normalised variance.

Outline We set up our programming language and smoothed semantics (Section 2) and prove it to be well-behaved (Section 3). In Section 4 we present our Diagonalisation Stochastic Gradient Descent procedure and its convergence. We review related work in Section 5 and conduct an experimental evaluation in Section 6. We conclude in Section 7.

2. Programming Language and Smoothing

Our starting point is a variant of the simply-typed λ-calculus (Barendregt et al., 2013) with reals, conditionals, and a non-standard *reparameterisable sampling* construct combining sampling from the standard normal and diffeomorphic transformations.

2.1. Syntax and Operational (Value) Semantics

We fix variables $\theta_1, \ldots, \theta_m$ of type R, which are the parameters to be optimised.

$$\begin{aligned}M ::= {} & x \mid \theta_i \mid \underline{r} \mid M + M \mid M \cdot M \mid \lambda x.\, M \mid M\, M \\ & \mid \mathbf{if}\ M < 0\ \mathbf{then}\ M\ \mathbf{else}\ M \\ & \mid \underline{\phi}_{\boldsymbol{\theta}}(M, \ldots, M, \mathbf{sample})\end{aligned}$$

where $r \in \mathbb{R}$ and $\phi : \mathbb{R}^m \times \mathbb{R}^\ell \to \mathbb{R}$ is a multivariate polynomial such that each $\phi_{\boldsymbol{\theta}}(y_1, \ldots, y_{\ell-1}, \cdot) : \mathbb{R} \to \mathbb{R}$ is a ***strong diffeomorphism***: i.e. for every *compact* $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$,

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \inf_{z_1,\ldots,z_\ell\in\mathbb{R}} \left|\frac{\partial \phi_{\boldsymbol{\theta}}}{\partial z_\ell}(z_1, \ldots, z_\ell)\right| > 0 \qquad (2)$$

In particular, every affine transformation is a strong diffeomorphism[1].

Intuitively, the latter non-standard construct samples from the standard normal distribution and immediately applies a transformation. In this way, for example, sampling from arbitrary normal distributions $\mathcal{N}(\mu, \sigma)$ can be simulated by $\underline{\phi}_{\mu,\sigma}(\mathbf{sample})$, where $\underline{\phi}_{\mu,\sigma}(z) \coloneqq \sigma \cdot z + \mu$.

In the spirit of (Kozen, 1979; Borgström et al., 2016), we view a probabilistic program as a deterministic function mapping parameters and *traces* (which are sequences of random samples) to values. For $\phi_\theta(z) \coloneqq z + \theta$ the function in Example 1.1 can be expressed as the value returned by the following term when substituting z as the sample:

$$\underline{-0.5} \cdot \theta^2 + (\mathbf{if}\ \underline{\phi}_\theta(\mathbf{sample}) < 0\ \mathbf{then}\ \underline{0}\ \mathbf{else}\ \underline{1})$$

Formally, we can endow the language with a standard simple type system and big-step operational semantics, keeping track of the trace of random samples (Borgström et al., 2016; Mak et al., 2021). In particular, for all $s \in \mathbb{R}$ (the samples)

$$\underline{\phi}_{\boldsymbol{\theta}}(\mathbf{sample}) \Downarrow^{(s)} \underline{\phi_{\boldsymbol{\theta}}(s)}$$

meaning, intuitively, that on trace s, the reparameterisable sampling construct $\underline{\phi}_{\boldsymbol{\theta}}(\mathbf{sample})$ evaluates to the numeral $\underline{\phi_{\boldsymbol{\theta}}(s)}$. Note that because our language does not have recursion/iteration there is a bound $n \in \mathbb{N}$ on the number of samples drawn on each path. Therefore, for all terms M of type R, values $r_1, \ldots, r_m \in \mathbb{R}$ (for the parameters $\theta_1, \ldots, \theta_m$) and samples $s_1, \ldots, s_n \in \mathbb{R}$, there exists a unique prefix s.t. $M[r_1/\theta_1, \ldots, r_m/\theta_m] \Downarrow^{(s_1,\ldots,s_\ell)} \underline{r}$; and we write $[\![M]\!]((r_1, \ldots, r_n), (s_1, \ldots, s_n)) = r$ in that case. In probabilistic programming, this function

$$[\![M]\!] : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$$

is called the ***value function*** of M (Borgström et al., 2016).

Henceforth, we do not distinguish between the *formal* parameters (variables) and the *actual* parameters (real numbers), and we write $[\![M]\!](\boldsymbol{\theta}, \mathbf{z})$ all the same, by abuse of notation.

Also note, that in the present work, because of the absence of a conditioning / score construct in our language, we are not explicitly concerned with the accompanying *weight* (or *density*) function (see (Borgström et al., 2016) for more details).

2.2. Problem Statement

We can phrase our optimisation problem as:

$$\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!](\boldsymbol{\theta}, \mathbf{z})] \qquad \textbf{(P)}$$

where $[\![M]\!] : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is the value-function of a term M of type R and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is the multivariate standard normal distribution.

For simplicity's sake we focus on expectations w.r.t. normal distributions in this work.

Since (only) piecewise polynomials are expressible in our language and (absolute) moments of normals are finite (cf. Lemma E.3) we obtain:

Proposition 2.1 (Integrability)**.** *For every* $\boldsymbol{\theta} \in \mathbb{R}^m$,

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[|[\![M]\!](\boldsymbol{\theta}, \mathbf{z})|] < \infty$$

Therefore, the optimisation problem (**P**) is well-defined.

Remark 2.2. As stated initially, we are mainly motivated by variational inference. If we restrict ourselves to normal distributions, each branch of the densities p and $q_{\boldsymbol{\theta}}$ are the product of normal densities, the logarithm of which are (bounded by) polynomials. In particular, the counterexample in (Lee et al., 2018, Proposition 2) is of this form.

The main advantage is that polynomials have pleasing (closure) properties. Note that permitting exponential functions injudiciously may render the optimisation problem (**P**) ill-defined, for example, $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\exp(z^2)] = \infty$ (independent of parameters).

2.3. Type System

We aim to ensure that the value functions of terms of type R are finite sums of functions $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of the form

$$(\boldsymbol{\theta}, \mathbf{z}) \mapsto \begin{cases} p(\boldsymbol{\theta}, \mathbf{z}) & \text{if for all } \psi \in \Psi_<,\ \psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) < 0, \\ & \text{and for all } \psi \in \Psi_\geq,\ \psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where

- (V1) $\Psi_<$ and $\Psi_\geq$ are finite sets of polynomials $\mathbb{R}^n \to \mathbb{R}$ which are not constant[2] 0
- (V2) $\boldsymbol{\phi}_{\boldsymbol{\theta}} = \langle \phi^{(1)}_{\boldsymbol{\theta}}, \ldots, \phi^{(n)}_{\boldsymbol{\theta}} \rangle : \mathbb{R}^n \to \mathbb{R}^n$ for strong diffeomorphisms (cf. Equation (2)) $\phi^{(i)}_{\boldsymbol{\theta}} : \mathbb{R}^i \to \mathbb{R}$.
- (V3) $p : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a polynomial.

Intuitively, $\Psi_<$ and $\Psi_\geq$ correspond to the constraints imposed by the if-branching, and $\boldsymbol{\phi}_{\boldsymbol{\theta}}$ is the transformation in a branch.

[1] A (contrived) counterexample is $\phi(z_1, z_2) \coloneqq z_1^2 z_2 + \frac{1}{3} z_1^2 z_2^3 - z_1 z_2^2 + z_2$ because $\frac{\partial \phi}{\partial z_2}(z_1, z_2) = z_1^2 + (z_1 z_2 - 1)^2$ has no positive lower bound despite being everywhere positive.

[2] This will be used in the uniform convergence proofs (cf. Proposition 3.3 and Lemma B.7).

An important consequence of this is that within a branch of computation (corresponding to a summand) all guards only depend on transformed variables, and they are obtained by the same diffeomorphism. This will allow us to cast partial derivatives w.r.t. a parameter θ in terms of gradients w.r.t. the noise variables $\mathbf{z}$ (see Equation (4)).

Remark 2.3. In particular, these conditions rule out optimisation problems whose objective functions are *discontinuous* such as[3]

$$\mathbb{E}_{z\sim\mathcal{N}(0,1)}[[\![\mathbf{if}\,\theta<0\,\mathbf{then}\,\underline{0}\,\mathbf{else}\,\underline{1}]\!](\theta,z)]=[\theta\geq 0]$$

where the expectation itself is discontinuous, rather than just the function inside the expectation.

Though important (in our view), it is unfortunate that general discontinuous *and* non-convex optimisation does not appear to have attracted much attention (in the machine learning community) with the notable exception of (Moreau and Aeyels, 2000; Zang, 1981; Bagirov et al., 2014; 2020).

Thus, our type system needs to enforce the constraint that guards do not directly depend on parameters θ_i (but, possibly, only indirectly through transformations). In particular,

$$\mathbf{if}\,\theta<0\,\mathbf{then}\,0\,\mathbf{else}\,1$$
$$(\lambda x.\,\mathbf{if}\,x<0\,\mathbf{then}\,0\,\mathbf{else}\,1)\,\theta$$

are not typable, but the following terms are typable

$$(\lambda x.\,\mathbf{if}\,x<0\,\mathbf{then}\,0\,\mathbf{else}\,1)\,(\underline{\phi_\theta}(\mathbf{sample}))$$
$$(\lambda x.\,\underline{-0.5}\cdot\theta^2+(\mathbf{if}\,x<0\,\mathbf{then}\,\underline{0}\,\mathbf{else}\,\underline{1}))\,(\underline{\phi_\theta}(\mathbf{sample}))$$

although for each sampled value the parameter θ may influence which branch is taken. Note that for $\phi_\theta(z)\coloneqq z+\theta$ the value-function of the latter term corresponds to Example 1.1.

To achieve this we use two kinds of typing judgements:

1. $\Gamma\vdash M:\tau$ for terms not involving parameters θ_i
2. $\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}} M:\tau$ for terms which may involve parameters, but not in guards. The variables in Γ can be used without any restriction, whilst variables of Δ may not be used in guards:

$$\frac{\Gamma\vdash L:R\quad\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}} M:\tau\quad\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}} N:\tau}{\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}}\mathbf{if}\,L<0\,\mathbf{then}\,M\,\mathbf{else}\,N:\tau}$$

Types are generated from the grammar: $\tau::=R\mid\tau\to\tau\mid\tau_{\boldsymbol{\theta}}$. Intuitively, terms of type $\tau_{\boldsymbol{\theta}}$ may depend on parameters. Thus their use must be restricted:

$$\frac{\Gamma\mid\Delta,y:\sigma\vdash_{\boldsymbol{\theta}} M:\tau}{\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}}\lambda y.\,M:\sigma_{\boldsymbol{\theta}}\to\tau}$$

[3]Given predicate ψ, we write $[\psi]$ as the Iverson bracket.

Besides, arguments to terms of type $\sigma\to\tau$ may not rely on parameters:

$$\frac{\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}} M:\sigma\to\tau\quad\Gamma\vdash M':\sigma}{\Gamma\mid\Delta\vdash_{\boldsymbol{\theta}} M\,M':\tau}$$

The full type systems are presented in Figure 4 in Appendix A.

2.4. Reparametrisation-Aware Symbolic Execution

Henceforth, we fix a typabable term $\emptyset\mid\emptyset\vdash_{\boldsymbol{\theta}} M:R$.

In order to show that the value function of a typable program is indeed the finite sum of functions of the form of Equation (3), we employ a *reparametrisation-aware symbolic execution*, which is a variant of the symbolic execution used by Mak et al. (2021). Intuitively, we

- collect constraints $\Psi_<$ and $\Psi_\geq$ due to branching;
- replace $\underline{\phi_{\boldsymbol{\theta}}}(P_1,\ldots,P_\ell,\mathbf{sample})$ with fresh sampling variable α_j and keep track of transformations $\boldsymbol{\phi_\theta}$.

Formally, we define a big-step-style symbolic execution (the full details are presented Appendix A.3)

$$M\Downarrow_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)} V$$

where M is a term of type R, V is a polynomial term over $\theta_1,\ldots,\theta_m$ and $\alpha_1,\ldots,\alpha_n$, the value function of which is a polynomial $\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$. This will ensure property (V3). Besides, as an invariant, properties (V1) and (V2) are maintained. Finally, this symbolic execution is *sound* and *complete* in the sense that

$$[\![M]\!](\boldsymbol{\theta},\mathbf{z})=\sum_{M\Downarrow_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)}P}[\![P]\!](\boldsymbol{\theta},\boldsymbol{\phi_\theta}(\mathbf{z}))\cdot\prod_{\psi\in\Psi_<}[\psi(\boldsymbol{\phi_\theta}(\mathbf{z}))<0]\cdot\prod_{\psi\in\Psi_\geq}[\psi(\boldsymbol{\phi_\theta}(\mathbf{z}))\geq 0]$$

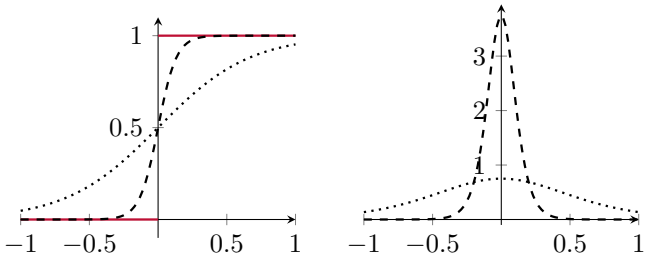
2.5. Smoothed Semantics

Finally, we define the smoothed value function $[\![M]\!]_k:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ parametrised by accuracy coefficient $k\in\mathbb{N}$ of a term $\emptyset\mid\emptyset\vdash_{\boldsymbol{\theta}} M:R$:

$$[\![M]\!]_k(\boldsymbol{\theta},\mathbf{z})\coloneqq\sum_{M\Downarrow_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)}P}[\![P]\!](\boldsymbol{\theta},\boldsymbol{\phi_\theta}(\mathbf{z}))\cdot\prod_{\psi\in\Psi_<}\sigma_k(-\psi(\boldsymbol{\phi_\theta}(\mathbf{z})))\cdot\prod_{\psi\in\Psi_\geq}\sigma_k(\psi(\boldsymbol{\phi_\theta}(\mathbf{z})))$$

where $\sigma_k(y)\coloneqq\sigma(\sqrt{k}\cdot y)$. Thus, we simply replace indicator functions by sigmoids (cf. Figure 2a).

We note that the smoothed value function can be computed efficiently by adapting (backward-mode) automatic differentiation (Kingma and Welling, 2014) to keep track of the



(a) σ_k (black, dashed/dotted) and the Heaviside step function (red, solid)

(b) Derivatives of σ_k

Figure 2. Sigmoid function σ_k and its derivative (dotted: $k = 9$, dashed: $k = 225$).

guards on branching. A formalisation of this would go beyond the scope of the present paper.

3. Properties of Smoothing

First, note that due to $0 \leq \sigma_k \leq 1$, $\mathbf{z} \mapsto |[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})|$ is bounded by the absolute value of a polynomial and therefore

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[|[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})|] < \infty$$

for each $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

3.1. Unbiasedness

Each $[\![M]\!]_k$ is clearly differentiable. The following is a consequence of a well-known result about exchanging differentiation and integration, which relies on the dominated convergence theorem (Klenke, 2014, Theorem 6.28).

Proposition 3.1 (Unbiasedness)**.** *For every $\boldsymbol{\theta} \in \mathbb{R}^m$ and $k \in \mathbb{N}$,*

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})] = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\nabla_{\boldsymbol{\theta}}[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})]$$

As a slight subtlety, we need to bound partial derivatives of $[\![M]\!]_k$ w.r.t. θ_i uniformly. Full details are presented in Appendix B.1.

As a consequence, once an accuracy coefficient $k \in \mathbb{N}$ is fixed, we can employ stochastic gradient descent to try to find stationary points of $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})]$.

3.2. Uniform Convergence

However, even if we can find a stationary point of the smoothed problem, *a priori* it is not obvious the extent to which this is an approximately stationary point of the original problem (**P**).

Henceforth, we assume that $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is compact.

Note that 0 is the discontinuity of the Heaviside function and $\sigma_k(0) = \frac{1}{2}$. Thus, we need to show that applications of sigmoids to 0 are "limited".

If $\psi : \mathbb{R}^n \to \mathbb{R}$ is a non-zero polynomial then the set of arguments which are approximately roots is "small[4]". Therefore, using the assumption that $\boldsymbol{\phi}_{\boldsymbol{\theta}}$ is a strong diffeomorphism, $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\sigma_k(\psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))]$ converges to $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) > 0]]$. In fact, the convergence is uniform for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (Lemma B.7).

From this and the fact that due to compactness of $\boldsymbol{\Theta}$, expectations $\mathbb{E}_{\mathbf{z}}[|p_{\boldsymbol{\theta}}(\mathbf{z})|]$ can be uniformly bounded for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ it follows relatively straightforwardly (Proposition C.1) that

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})] \xrightarrow{\text{unif}} \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!](\boldsymbol{\theta}, \mathbf{z})]$$

as $k \to \infty$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proving uniform convergence of gradients is more challenging. The main difficulty is that $\lim_{k\to\infty} \sigma_k'(0) = \infty$ (cf. Figure 2b). Thus, there exist $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{z} \in \mathbb{R}^n$ for which $\nabla_{\boldsymbol{\theta}}[\![M]\!]_k(\boldsymbol{\theta}, \mathbf{z})$ is unbounded:

$$\nabla_\theta[\![\mathbf{if}\ \underline{\phi}_\theta(z) < 0\ \mathbf{then}\ \underline{0}\ \mathbf{else}\ \underline{1}]\!]_k(\theta, z) = \sigma_k'(z + \theta) \to \infty$$

whenever $\theta = -z$ for $\phi_\theta(z) = z + \theta$.

Fortunately, this effect again only occurs at "small" subsets of the latent space. The key technical trick to taming this behaviour is to cast partial derivatives w.r.t. the parameters in terms of gradients w.r.t. the latent variables and to use integration by parts, thus eliminating derivatives of σ_k.

Example 3.2. Consider the (typable) term M defined by

$$(\lambda x.\,(\mathbf{if}\ x < 0\ \mathbf{then}\ \underline{0}\ \mathbf{else}\ (\mathbf{if}\ x - \underline{1} < 0\ \mathbf{then}\ \underline{1}\ \mathbf{else}\ \underline{0}))\ (\underline{\phi}_\theta(\mathbf{sample}))$$

where $\phi_\theta(z) \coloneqq c \cdot z + \theta$ and $0 \neq c \in \mathbb{R}$ is a constant. Define

$$\Psi_k(\alpha) \coloneqq \sigma_k(\alpha) \cdot \sigma_k(1 - \alpha)$$

and note that $[\![M]\!]_k(\theta, z) = \Psi_k(\phi_\theta(z))$. By the chain rule we obtain:

$$\frac{\partial(\Psi_k \circ \phi_{(-)})}{\partial\theta}(\theta, z) = \frac{1}{c} \cdot \frac{\partial(\Psi_k \circ \phi_{(-)})}{\partial z}(\theta, z) \quad (4)$$

Note that for this it is essential that the guards, i.e. which correspond to the factors in the definition of Ψ_k, do not depend on the parameters, but that they are composed with the same transformation[5] $\boldsymbol{\phi}_{\boldsymbol{\theta}}$.

[4] Recall that the set of roots of polynomials is negligible (Caron, 2005).

[5] in this particular branch of the computation

This enables us to employ integration by parts:

$$\begin{aligned}&\mathbb{E}_z\left[\frac{\partial(\Psi_k\circ\phi_{(-)})}{\partial\theta}(\theta,z)\right]\\&=\int\mathcal{N}(z)\cdot\frac{1}{c}\cdot\frac{\partial(\Psi_k\circ\phi_{(-)})}{\partial z}(\theta,z)\,\mathrm{d}z\\&=\frac{1}{c}\left([\mathcal{N}(z)\cdot\Psi_k(\phi_\theta(z))]_{-\infty}^{\infty}+\mathbb{E}_z[z\cdot\Psi_k(\phi_\theta(z))]\right)\end{aligned}$$

since $\frac{\partial\mathcal{N}}{\partial x}(x)=-x\cdot\mathcal{N}(x)$. Due to $0\leq\Psi_k(\phi_\theta(z))\leq 1$ the first summand is 0 and the uniform convergence of the second follows as above.

The reasoning can be generalised by employing inverse Jacobians in Equation (4). For this we exploit the fact that the diffeomorphisms are strong in the sense of Equation (2).

Proposition 3.3 (Uniform Convergence of Gradients)**.**

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!]_k(\boldsymbol{\theta},\mathbf{z})]\xrightarrow{\text{unif}}\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!](\boldsymbol{\theta},\mathbf{z})]$$

as $k\to\infty$ *for* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

The following example illustrates that it is *not* sufficient to restrict terms $\underline{\phi}_{\boldsymbol{\theta}}(M_1,\ldots,M_\ell,\mathbf{sample})$ to polynomial *bijections*:

Example 3.4 (Divergence)**.** Suppose $M\equiv$ **if** $\underline{\phi}_\theta(\mathbf{sample})<0$ **then** $\underline{0}$ **else** $\underline{1}$, where $\phi_\theta(z):=z^3+\theta$. Note that $\phi_\theta:\mathbb{R}\to\mathbb{R}$ is not a diffeomorphism because $\phi_\theta^{-1}(\alpha)=\sqrt[3]{\alpha-\theta}$ is not differentiable at $\alpha=\theta$. Then

$$\begin{aligned}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[[\![M]\!](\theta,z)]&=\int_{-\sqrt[3]{-\theta}}^{\infty}\mathcal{N}(z\mid 0,1)\,\mathrm{d}z\\\frac{\partial}{\partial\theta}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[[\![M]\!](\theta,z)]&=\frac{1}{3}\cdot\mathcal{N}(-\sqrt[3]{-\theta}\mid 0,1)\cdot\theta^{-\frac{2}{3}}\end{aligned}$$

Therefore $\theta\mapsto\mathbb{E}_{z\sim\mathcal{N}(0,1)}[[\![M]\!](\boldsymbol{\theta},z)]$ is not differentiable at 0. Besides, for $\theta=0$,

$$\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\frac{\partial}{\partial\theta}[[\![M]\!]_k(\boldsymbol{\theta},z)\right]=\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma_k'(z^3)\right]\to\infty$$

The significance of Proposition 3.3 is that it provides the theoretical basis for obtaining approximately stationary points for the original problem (**P**): given an error tolerance $\epsilon>0$ there exists an accuracy coefficient k such that $\|\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}}[[\![M]\!]_k(\boldsymbol{\theta},\mathbf{z})]-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}}[[\![M]\!](\boldsymbol{\theta},\mathbf{z})]\|<\epsilon$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Thus, a stationary point $\boldsymbol{\theta}^*\in\boldsymbol{\Theta}$ of the smoothed problem for accuracy coefficient k (which may be obtained by SGD) is approximately stationary:

$$\|\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\![M]\!]_k(\boldsymbol{\theta},\mathbf{z})]\|<\epsilon$$

4. Diagonalisation SGD

So far, we have investigated the unbiasedness of smoothed gradient estimators and their use in gradient-based optimisation methods (conditional on their convergence to stationary points) for a *fixed* accuracy coefficient.

The objective of this section is twofold: Firstly, to obtain an (unconditional) convergence result. Secondly, not to *fix* accuracy coefficient but rather enhance it *during* the optimisation.

To formalise this, suppose for each $k\in\mathbb{N}$, $f_k:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is differentiable. We define a ***Diagonalisation Stochastic Gradient Descent*** sequence:

$$\boldsymbol{\theta}_{k+1}:=\boldsymbol{\theta}_k-\gamma_k\nabla_{\boldsymbol{\theta}}f_k(\boldsymbol{\theta}_k,\mathbf{z}_k)\qquad\mathbf{z}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$$

The qualifier "diagonal" highlights that, in contrast to standard stochastic gradient descent, we are not using the gradient of the same function f for each step $\boldsymbol{\theta}_k$, but rather we are using the gradient of f_k.

The following exploits Taylor's theorem and can be obtained by minor tweaks to standard convergence proofs of (stochastic) gradient descent (see e.g. (Bertsekas and Tsitsiklis, 2000) or (Bertsekas, 2015, Chapter 2)):

Proposition 4.1 (Convergence)**.** *Suppose* $\gamma_k\in\Theta(1/k)$ *and* $g_k(\boldsymbol{\theta}):=\mathbb{E}_{\mathbf{z}}[f_k(\boldsymbol{\theta},\mathbf{z})]$ *and* $g(\boldsymbol{\theta}):=\mathbb{E}_{\mathbf{z}}[f(\boldsymbol{\theta},\mathbf{z})]$ *are well-defined and differentiable.*

Suppose there exist $\{\boldsymbol{\theta}_k\mid k\in\mathbb{N}\}\subseteq\boldsymbol{\Theta}\subseteq\mathbb{R}^m$, $L>0$ *and* $\epsilon>0$ *s.t. for all* $k\in\mathbb{N}$ *and* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$,

(D1) $\nabla_{\boldsymbol{\theta}}g_k(\boldsymbol{\theta})=\mathbb{E}_{\mathbf{z}}[\nabla_{\boldsymbol{\theta}}f_k(\boldsymbol{\theta},\mathbf{z})]$

(D2) $|g_{k+1}(\boldsymbol{\theta})-g_k(\boldsymbol{\theta})|<k^{-1-\epsilon}\cdot L$

(D3) $\|\nabla g_k(\boldsymbol{\theta})-\nabla g(\boldsymbol{\theta})\|^2<k^{-\epsilon}\cdot L$

(D4) $\mathbb{E}_{\mathbf{z}}[\|\nabla_{\boldsymbol{\theta}}f_k(\boldsymbol{\theta},\mathbf{z})\|^2]<L$

(D5) $\|\mathbf{H}\,g_k(\boldsymbol{\theta})\|<L$

Then $\inf_{i\in\mathbb{N}}\mathbb{E}[\|\nabla g(\boldsymbol{\theta}_i)\|^2]=0$.

Intuitively, (D1) is the unbiasedness of the Monte Carlo estimator, (D2) and (D3) stipulate uniform convergence of the g_k and their gradient, respectively, and (D4) and (D5) guarantee a uniform bound on the "variance" and the Hessian, respectively.

4.1. Application to Smoothing

Finally, we would like to instantiate $f_k:=[[\![M]\!]_k$ and $f:=[[\![M]\!]$, the (smoothed) value function of a typable term. Essentially, we have already convinced ourselves that condition (D1) to (D3) are satisfied (see also Propositions C.1, B.4 and B.13).

(D4) and (D5) can be verified using similar tricks as in Proposition 3.3, in particular Equation (4) enabling integration by parts (Propositions C.2 and C.3).

Thus, we obtain our main result:

Theorem 4.2 (Convergence on Typable Programs)**.** *If* $\emptyset \mid \emptyset \vdash_{\boldsymbol{\theta}} M : R$ *then a DSGD sequence* $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}}$ *is unbounded or has a* stationary *accumulation point for the optimisation problem* (**P**)*.*

5. Related Work

(Lee et al., 2018) is both the starting point for our work and the most natural source for comparison. They correct the (biased) reparametrisation gradient estimator for non-differentiable models by additional non-trivial *boundary* terms. They present an efficient method for *affine* guards only. Besides, they are not concerned with the *convergence* of gradient-based optimisation procedures; nor do they discuss how assumptions they make may be manifested in a programming language.

In the context of the reparametrisation gradient, Maddison et al. (2017) and Jang et al. (2017) relax discrete random variables in a continuous way, effectively dealing with a specific class of discontinuous models. Zang (1981) use a similar smoothing for discontinuous optimisation but they do not consider a full programming language.

Lew et al. (2020) and Wang et al. (2021) also discuss type systems in the context of variational inference for ensuring *absolute continuity*, a necessary (but not sufficient) property avoided by our formulation of the optimisation problem (**P**).

6. Experimental Evaluation

We evaluate our smoothed gradient estimator (SMOOTH) against the biased reparameterisation estimator (REPARAM), the unbiased correction of it (LYY18) due to (Lee et al., 2018), and the unbiased (SCORE) estimator. The implementation is written in Python using the `jax` library to provide automatic differentiation which is used to implement each of the above estimators for an arbitrary probabilistic program.

We focus on the three models `temperature`, `textmsg` and `influenza` as described in (Lee et al., 2018). Each of these has multiple non-differentiablities arising as follows:

- `temperature` (Soudjani et al., 2017) models a controller keeping the temperature of a room within set bounds. The discontinuity arises from the discrete state of the controller, being either on or off. The model has a 41-dimensional latent variable and 80 if-statements.
- `textmsg` (Davidson-Pilon, 2015) models daily text message rates and the goal is to discover a change in the rate over the 74-day period of data given. The non-differentiability arises from the point at which the rate is modelled to change. The model has a 3-dimensional latent variable and 37 if-statements.
- `influenza` (Shumway and Stoffer, 2005) models the US influenza mortality for 1969. In each month, the rate depends on the dominant virus strain being of type 1 or type 2 which produces a non-differentiablity for each month. The model has a 37-dimensional latent variable and 24 if-statements.

For each model, we follow a similar experimental setup to that of Lee et al. (2018) by optimising the ELBO function using each gradient estimator. We use the stochastic optimiser `Adam` with a step size of `0.001` for 10000 iterations; and, for each iteration, we use a number $N \in \{1, 8, 16\}$ of Monte Carlo samples from the chosen estimator to compute the gradient. As in (Lee et al., 2018), the LYY18 estimator doesn't compute the full boundary surface term but takes a single subsample for each estimate. For the SMOOTH estimator, for iteration $k \in \mathbb{N}$ we use the slightly modified $\sigma_k(x) \coloneqq \sigma(10\sqrt{k} \cdot x)$.

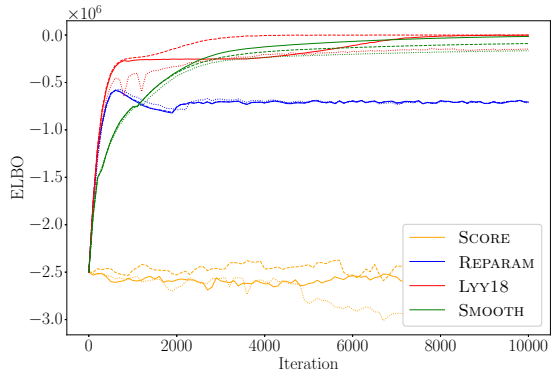
For every 100 iterations, we take 1000 samples of the estimator to estimate the current ELBO value and the variance of the gradient. Since the gradient is a vector, the variance is taken in two ways: averaging the component-wise variances and the variance of the L2 norm.

We separately benchmark each estimator by computing the number of iterations each can complete in a fixed time budget; the computational cost of each estimator is then estimated to be the reciprocal of this number. This then allows us to compute a set of *work-normalised* variances (Botev and Ridder, 2017) for each estimator which is taken to be the product of the computational cost and the variance.

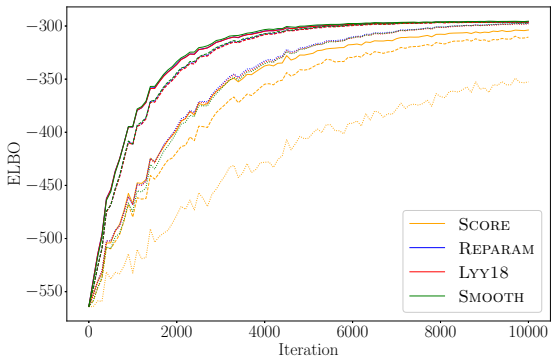
Table 1 presents the ratios of the work-normalised variances for each estimator with respect to the SCORE estimator. The ELBO trajectories are displayed in Figure 3.

The trajectories all show the SMOOTH estimator performing comparably to the LYY18 estimator. (Notice that, in the `temperature` model, the REPARAM estimator is biased.) The benchmark and variance results are also promising, with the SMOOTH estimator having the lowest work-normalised variance across the board. This is partly because our estimator has a comparable computational cost to the REPARAM estimator, but also because it has a low variance, likely from differentiability of the smoothed models.

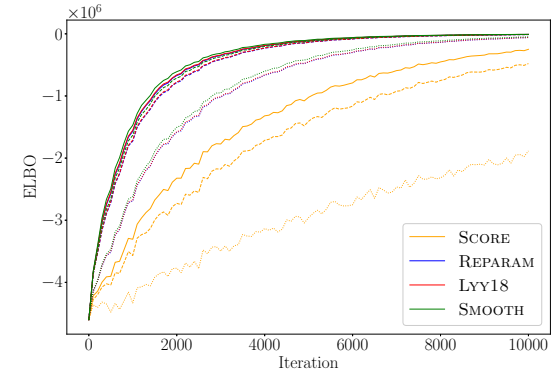
Finally, our `jax` implementation of the LYY18 estimator followed the inventors' implementation as closely as possible so we claim it is a fair comparison point. To support our claim, in the supplementary material we include results from adding our SMOOTH estimator to the code made publicly available by Lee et al. (2018) and show that the results are still comparable.



(a) `temperature`



(b) `textmsg`



(c) `influenza`

Figure 3. ELBO trajectories for each model. A single colour is used for each estimator and the batch size $N = 1, 8, 16$ is represented by dotted, dashed and solid lines respectively.

Table 1. Computational cost and work-normalised variances, all given as ratios with respect to the SCORE estimator (omitted since it would be all 1s).

(a) `temperature`

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	1.28e+00	1.13e-08	1.56e-08
LYY18	8.89e+00	7.98e-09	1.78e-07
SMOOTH	2.13e+00	2.51e-11	1.04e-10

(b) `textmsg`

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	8.83e+00	1.28e-02	1.03e-02
LYY18	1.07e+01	2.34e-02	2.30e-02
SMOOTH	1.11e+01	1.16e-02	9.59e-03

(c) `influenza`

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	1.64e+00	8.64e-04	3.77e-04
LYY18	8.44e+00	4.45e-03	1.96e-03
SMOOTH	1.73e+00	8.19e-04	3.74e-04

7. Concluding Remarks

We have proposed a variant of SGD, *Diagonalisation Stochastic Gradient Descent*, and shown *provable* convergence. Our approach is based on a smoothed interpretation of (possibly) discontinuous programs, which also yields unbiased gradient estimators. We have designed a type system to impose a mild restriction on the programming language. Whilst this provides sufficient conditions for theoretical guarantees, we stress that DSGD and the smoothed unbiased gradient estimator can even be applied to programs which are *not* typable.

Experiments with our prototype implementation confirm the benefits of reduced variance and unbiasedness. Compared to the unbiased correction of the reparametrised gradient estimator, our estimator has a similar convergence, but is simpler, faster, and attains orders of magnitude (2 to 3,000 x) reduction in work-normalised variance.

Future Directions We aim to extend the programming language beyond piecewise polynomials whilst still retaining the pleasing properties derived here. We anticipate this will mostly require making the type system more elaborate to enforce properties which hold naturally for polynomials. In that respect the present language is at a sweet spot. Another avenue for future research is to support recursion in the programming language. In a different direction, we expect the smoothing-cum-Diagonalisation-SGD approach to yield efficient gradient estimators for normalising flows with discrete distributions.

References

Adil Bagirov, Napsu Karmitsa, and Marko M. Mäkelä. *Introduction to Nonsmooth Optimization: Theory, Practice and Software*. Springer, 2014.

Adil M. Bagirov, Manlio Gaudioso, Napsu Karmitsa, Marko M. Mäkelä, and Sona Taheri. *Numerical Nonsmooth Optimization: State of the Art Algorithms*. Springer, 2020.

Hendrik Pieter Barendregt, Wil Dekkers, and Richard Statman. *Lambda Calculus with Types*. Perspectives in logic. Cambridge University Press, 2013.

Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642, 2000.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi: 10.1137/141000671.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

Johannes Borgström, Ugo Dal Lago, Andrew D. Gordon, and Marcin Szymczak. A lambda-calculus foundation for universal probabilistic programming. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, ICFP 2016, Nara, Japan, September 18-22, 2016*, pages 33–46, 2016.

Zdravko Botev and Ad Ridder. Variance Reduction. In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. 2017.

Richard Caron. The zero set of a polynomial. Technical Report WMSR 05-03, University of Windsor, Department of Mathematics and Statistics, May 2005.

Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

C. Davidson-Pilon. *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley Professional, 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer London, 2014.

Dexter Kozen. Semantics of probabilistic programs. In *20th Annual Symposium on Foundations of Computer Science, San Juan, Puerto Rico, 29-31 October 1979*, pages 101–114, 1979.

Wonyeol Lee, Hangyeol Yu, and Hongseok Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5558–5568, 2018.

Alexander K. Lew, Marco F. Cusumano Towner, Benjamin Sherman, Michael Carbin, and Vikash K. Mansinghka. A type system and semantics for sound programmable inference in probabilistic languages. *PACMPL*, 4(POPL), 2020.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Carol Mak, C.-H. Luke Ong, Hugo Paquet, and Dominik Wagner. Densities of almost surely terminating probabilistic programs are differentiable almost everywhere. In Nobuko Yoshida, editor, *Programming Languages and Systems - 30th European Symposium on Programming, ESOP 2021, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2021, Luxembourg City, Luxembourg, March 27 - April 1, 2021, Proceedings*, volume 12648 of *Lecture Notes in Computer Science*, pages 432–461. Springer, 2021.

Luc Moreau and Dirk Aeyels. Optimization of discontinuous functions: a generalised theory of differentiation. *SIAM J. OPTIM.*, 11(1):53–69, 2000.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence*

and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014, pages 814–822, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Walter Rudin. *Principles of mathematical analysis, 3rd edition*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1973.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer-Verlag, 2005.

Sadegh Esmaeil Zadeh Soudjani, Rupak Majumdar, and Tigran Nagapetyan. Multilevel monte carlo method for statistical model checking of hybrid systems. In Nathalie Bertrand and Luca Bortolussi, editors, *Quantitative Evaluation of Systems - 14th International Conference, QEST 2017, Berlin, Germany, September 5-7, 2017, Proceedings*, volume 10503 of *Lecture Notes in Computer Science*, pages 351–367. Springer, 2017.

Michalis K. Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1971–1979, 2014.

Di Wang, Jan Hoffmann, and Thomas W. Reps. Sound probabilistic inference via guide types. In Stephen N. Freund and Eran Yahav, editors, *PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 20211*, pages 788–803. ACM, 2021.

Israel Zang. Discontinuous optimization by smoothing. *Mathematics of Operations Research*, 6(1):140–152, 1981. ISSN 0364765X, 15265471.

A. Supplementary Materials for Section 2

A.1. Supplementary Materials for Section 2.1

We recall a big-step operational semantics in the style of (Borgström et al., 2016; Mak et al., 2021). $M \Downarrow^{\mathbf{s}} V$ means that the closed term M reduces to value V using the random samples $\mathbf{s}$. A value is either a real constants $\underline{r}$ or a λ-abstraction $\lambda x.\, M$.

$$\frac{}{V \Downarrow^{()} V} \qquad \frac{M \Downarrow^{\mathbf{s}} \underline{r} \quad M' \Downarrow^{\mathbf{s}'} \underline{r'}}{M \circ M' \Downarrow^{\mathbf{s} +\!\!+ \mathbf{s}'} \underline{r \oplus r'}}\ \circ \in \{+, \cdot\}$$

$$\frac{M_1 \Downarrow^{\mathbf{s}^{(1)}} \underline{r_1} \quad \cdots \quad M_\ell \Downarrow^{\mathbf{s}^{(\ell)}} \underline{r_\ell}}{\underline{\phi}_{\boldsymbol{\theta}}(M_1, \ldots, M_\ell, \mathbf{sample}) \Downarrow^{\mathbf{s}^{(1)} +\!\!+ \cdots +\!\!+ \mathbf{s}^{(\ell)} +\!\!+ (s)} \underline{\phi_{\boldsymbol{\theta}}(r_1, \ldots, r_\ell, s)}}\ s \in \mathbb{R}$$

$$\frac{M \Downarrow^{\mathbf{s}} \lambda y.\, N \quad M' \Downarrow^{\mathbf{s}'} V \quad N[V/y] \Downarrow^{\mathbf{s}''} V'}{M\, M' \Downarrow^{\mathbf{s} +\!\!+ \mathbf{s}' +\!\!+ \mathbf{s}''} V'}$$

$$\frac{L \Downarrow^{\mathbf{s}} \underline{r} \quad M \Downarrow^{\mathbf{s}'} V}{\mathbf{if}\ L < 0\ \mathbf{then}\ M\ \mathbf{else}\ N \Downarrow^{\mathbf{s} +\!\!+ \mathbf{s}'} V}\ r < 0 \qquad \frac{L \Downarrow^{\mathbf{s}} \underline{r} \quad N \Downarrow^{\mathbf{s}'} V}{\mathbf{if}\ L < 0\ \mathbf{then}\ M\ \mathbf{else}\ N \Downarrow^{\mathbf{s} +\!\!+ \mathbf{s}'} V}\ r \geq 0$$

where $(s_1, \ldots, s_n) +\!\!+ (s_{n+1}, \ldots, s_{n'}) \coloneqq (s_1, \ldots, s_{n'})$ concatenates tuples.

A.2. Supplementary Materials for Section 2.3

The type systems are presented in Figure 4.

Lemma A.1 (Weakening). *1. If $\Gamma \vdash M : \tau$ and $\Gamma \subseteq \Gamma'$ then $\Gamma' \vdash M : \tau$.*

2. If $\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau$, $\Gamma \subseteq \Gamma'$ and $\Delta \subseteq \Delta'$ then $\Gamma' \mid \Delta' \vdash_{\boldsymbol{\theta}} M : \tau$.

Lemma A.2 (Substitution). *1. If $\Gamma, x : \sigma \vdash M : \tau$ and $\Gamma \vdash N : \sigma$ then $\Gamma \vdash M[N/x] : \tau$.*

2. If $\Gamma, x : \sigma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau$ and $\Gamma \vdash N : \sigma$ then $\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M[N/x] : \tau$.

3. If $\Gamma \mid \Delta, x : \sigma \vdash_{\boldsymbol{\theta}} M : \tau$ and $\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} N : \sigma$ then $\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M[N/x] : \tau$.

A.3. Supplementary Materials for Section 2.4

First-order polynomial terms (over $\theta_1, \ldots, \theta_m$ and $\alpha_1, \ldots, \alpha_n$) and symbolic values can be characterised as:

$$P ::= \underline{r} \mid \theta_j \mid \alpha_i \mid P + P \mid P \cdot P$$
$$V ::= P \mid \lambda y.\, M$$

Note that symbolic values (typically denoted by V) are different from the ones used for the operational semantics in Appendix A.1. However, no confusion should arise because henceforth we are only referring to symbolic values.

The following is obvious:

Lemma A.3. *1. If $\alpha_1, \ldots, \alpha_n \vdash P : R$ then $[\![P]\!] : \mathbb{R}^n \to \mathbb{R}$ is a polynomial.*

2. If $\alpha_1, \ldots, \alpha_n \mid \emptyset \vdash_{\boldsymbol{\theta}} P : R$ then $[\![P]\!] : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a polynomial.

Similarly as in (Mak et al., 2021), we introduce a (big-step-style) *symbolic execution* $n, n' \rhd M \Downarrow_{\boldsymbol{\phi}}^{(\Psi_<, \Psi_\geq)}$ to closely mirror the operational semantics recalled in Appendix A.1. Intuitively, terms $\underline{\phi}_{\boldsymbol{\theta}}(P_1, \ldots, P_\ell, \mathbf{sample})$ are replaced with fresh sampling variable $\alpha_{n+1}, \ldots, \alpha_{n'}$ and $\boldsymbol{\phi}_{(-)} = \langle \phi_{(-)}^{(n+1)}, \ldots, \phi_{(-)}^{n'} \rangle$ keeps track of the transformations. Besides, $\Psi_<$ and $\Psi_\geq$

$$\frac{}{\Gamma \mid \Delta \vdash \underline{r} : \tau}\ r \in \mathbb{R} \qquad \frac{}{\Gamma, x : \tau \vdash x : \tau}$$

$$\frac{\Gamma \vdash M : R \quad \Gamma \vdash M' : R}{\Gamma \vdash M \circ M' : R}\ \circ \in \{+, \cdot\} \qquad \frac{\Gamma \vdash M_1 : R \quad \cdots \quad \Gamma \vdash M_\ell : R}{\Gamma \vdash \underline{\phi}_{\boldsymbol{\theta}}(M_1, \ldots, M_\ell, \mathbf{sample}) : R}$$

$$\frac{\Gamma \vdash L : R \quad \Gamma \vdash M : \tau \quad \Gamma \vdash N : \tau}{\Gamma \vdash \mathbf{if}\ L < 0\ \mathbf{then}\ M\ \mathbf{else}\ N : \tau}$$

$$\frac{\Gamma, y : \sigma \vdash M : \tau}{\Gamma \vdash \lambda y.\, M : \sigma \to \tau} \qquad \frac{\Gamma \vdash M : \sigma \to \tau \quad \Gamma \vdash M' : \sigma}{\Gamma \vdash M\, M' : \tau}$$

(a) Type systems for terms not involving parameters θ_i

$$\frac{}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \underline{r} : \tau}\ r \in \mathbb{R} \qquad \frac{}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \theta_i : R} \qquad \frac{}{\Gamma, x : \tau \mid \Delta \vdash_{\boldsymbol{\theta}} x : \tau} \qquad \frac{}{\Gamma \mid \Delta, x : \tau \vdash_{\boldsymbol{\theta}} x : \tau}$$

$$\frac{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : R \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M' : R}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M \circ M' : R}\ \circ \in \{+, \cdot\} \qquad \frac{\Gamma \vdash M_1 : R \quad \cdots \quad \Gamma \vdash M_\ell : R}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \underline{\phi}_{\boldsymbol{\theta}}(M_1, \ldots, M_\ell, \mathbf{sample}) : R}$$

$$\frac{\Gamma \vdash L : R \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} N : \tau}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \mathbf{if}\ L < 0\ \mathbf{then}\ M\ \mathbf{else}\ N : \tau}$$

$$\frac{\Gamma, y : \sigma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \lambda y.\, M : \sigma \to \tau} \qquad \frac{\Gamma \mid \Delta, y : \sigma \vdash_{\boldsymbol{\theta}} M : \tau}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \lambda y.\, M : \sigma_{\boldsymbol{\theta}} \to \tau}$$

$$\frac{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \sigma \to \tau \quad \Gamma \vdash M' : \sigma}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M\, M' : \tau} \qquad \frac{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \sigma_{\boldsymbol{\theta}} \to \tau \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M' : \sigma}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M\, M' : \tau}$$

(b) Type system enforcing restriction on the dependence of guards on parameters

Figure 4. Type systems

collect the conditions on the variables due to branching.

$$\frac{}{n, n \rhd V \Downarrow^{(\emptyset,\emptyset)} V}$$

$$\frac{n, n' \rhd M \Downarrow_{\boldsymbol{\phi}}^{\Psi} P \quad n', n'' \rhd M' \Downarrow_{\boldsymbol{\phi}'}^{\Psi'} P'}{n, n'' \rhd M \circ M' \Downarrow_{\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle}^{\Psi \cup \Psi'} P \circ P'} \circ \in \{+, \cdot\}$$

$$\frac{n, n_1 \rhd M_1 \Downarrow_{\boldsymbol{\phi}^{(1)}}^{\Psi_1} P_1 \quad \cdots \quad n_{\ell-1}, n_\ell \rhd M_\ell \Downarrow_{\boldsymbol{\phi}^{(\ell)}}^{\Psi_\ell} P_\ell}{n, n_\ell + 1 \rhd \underline{\phi}_{\boldsymbol{\theta}}(M_1, \ldots, M_\ell, \mathbf{sample}) \Downarrow_{\langle \boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(\ell)}, \phi^{(n_\ell+1)} \rangle}^{\Psi_1 \cup \cdots \cup \Psi_\ell} \alpha_{n_\ell+1}}$$

$$\frac{n, n' \rhd M \Downarrow_{\boldsymbol{\phi}}^{\Psi} \lambda y.\, N \quad n', n'' \rhd M' \Downarrow_{\boldsymbol{\phi}'}^{\Psi'} V \quad n'', n''' \rhd N[V/y] \Downarrow_{\boldsymbol{\phi}''}^{\Psi''} V'}{n, n''' \rhd M\, M' \Downarrow_{\langle \boldsymbol{\phi}, \boldsymbol{\phi}', \boldsymbol{\phi}'' \rangle}^{\Psi \cup \Psi' \cup \Psi''} V'}$$

$$\frac{n, n' \rhd L \Downarrow_{\boldsymbol{\phi}}^{\Psi} P \quad n', n'' \rhd M \Downarrow_{\boldsymbol{\phi}'}^{\Psi'} V}{n, n'' \rhd \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N \Downarrow_{\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle}^{\Psi \cup \Psi' \cup (\{[\![P]\!]\}, \emptyset)} V} [\![P]\!] \neq 0 \qquad \frac{n, n' \rhd L \Downarrow_{\boldsymbol{\phi}}^{\Psi} P \quad n', n'' \rhd N \Downarrow_{\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle}^{\Psi'} V}{n, n'' \rhd \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N \Downarrow_{\boldsymbol{\phi}, \boldsymbol{\phi}'}^{\Psi \cup \Psi' \cup (\emptyset, \{[\![P]\!]\})} V} [\![P]\!] \neq 0$$

$$\frac{n, n' \rhd L \Downarrow_{\boldsymbol{\phi}}^{\Psi} P \quad n', n'' \rhd N \Downarrow_{\boldsymbol{\phi}'}^{\Psi'} V}{n, n'' \rhd \mathbf{if}\, L < 0\, \mathbf{then}\, M\, \mathbf{else}\, N \Downarrow_{\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle}^{\Psi \cup \Psi'} V} [\![P]\!] = 0$$

We silently lift the domains of functions (e.g. from $\mathbb{R}^{n'}$ to $\mathbb{R}^n$ when $n' \leq n$) and use the shorthands

- $\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle \coloneqq \langle \phi_{(-)}^{(n)}, \ldots, \phi_{(-)}^{(n'')} \rangle$, for $\boldsymbol{\phi} = \langle \phi_{(-)}^{(n)}, \ldots, \phi_{(-)}^{(n')} \rangle$ and $\boldsymbol{\phi}' = \langle \phi_{(-)}^{(n'+1)}, \ldots, \phi_{(-)}^{(n'')} \rangle$ and
- $(\Psi_<, \Psi_\geq) \cup (\Psi'_<, \Psi'_\geq) \coloneqq (\Psi_< \cup \Psi'_<, \Psi_\geq \cup \Psi'_\geq)$.

Besides, in the rule for the reparameterisable sampling construct,

$$\phi_{\boldsymbol{\theta}}^{(n_\ell+1)} \coloneqq \phi_{\boldsymbol{\theta}} \circ \langle [\![P_1]\!], \ldots, [\![P_\ell]\!], \pi_{n_\ell+1} \rangle) \circ \langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)}, \ldots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)}, \pi_{n_{\ell+1}} \rangle$$

Note that in case the guard evaluates to the trivial constantly 0 polynomial, we ignore the if-branch and do not record the constraint in $\Psi_\geq$. This ensures that we can prove the following invariant, which is important for the uniform convergence proofs (cf. Lemma B.7):

Proposition A.4 (Invariant). *If $\alpha_1, \ldots, \alpha_n \mid \emptyset \vdash_{\boldsymbol{\theta}} M : \tau$ and $n', n'' \rhd M \Downarrow_{\boldsymbol{\phi}}^{(\Psi_<, \Psi_\geq)} V$, where $n \leq n'$ then*

1. $n \leq n' \leq n''$
2. $\alpha_1, \ldots, \alpha_{n''} \mid [] \vdash_{\boldsymbol{\theta}} V : \tau$
3. *$\boldsymbol{\phi} = \langle \phi^{(n'+1)}, \ldots, \phi^{(n'')} \rangle$ and for each $n' < j \leq n''$ and $\boldsymbol{\theta} \in \mathbb{R}^m$, $\phi_{(-)}^{(j)} : \mathbb{R}^m \times \mathbb{R}^j \to \mathbb{R}$ is a strong polynomial diffeomorphism (cf. Equation (2)).*
4. *$\Psi_<$ and $\Psi_\geq$ are sets of polynomials $\mathbb{R}^{n''} \to \mathbb{R}$ which are not constant 0*

Proof sketch. Fairly routine induction. The case for transformed samples uses Lemma A.3 and the fact that

$$\begin{aligned}
\frac{\partial \phi_{\boldsymbol{\theta}}^{(n_\ell+1)}}{\partial z_{n_\ell+1}}(\boldsymbol{\theta},\mathbf{z}) &= \mathbf{J}(\phi_{\boldsymbol{\theta}} \circ \langle [\![P_1]\!],\dots,[\![P_\ell]\!],\pi_{n_{\ell+1}}\rangle))(\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\mathbf{z},\boldsymbol{\theta})) \cdot J_{z_{n_\ell+1}}\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\boldsymbol{\theta},\mathbf{z}) \\
&= \mathbf{J}(\phi_{\boldsymbol{\theta}} \circ \langle [\![P_1]\!],\dots,[\![P_\ell]\!],\pi_{n_{\ell+1}}\rangle))(\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\mathbf{z},\boldsymbol{\theta})) \cdot \mathbf{e}_{n_\ell+1} \\
&= (\nabla\phi(\langle [\![P_1]\!],\dots,[\![P_\ell]\!],\pi_{n_{\ell+1}}\rangle(\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\mathbf{z},\boldsymbol{\theta}))))^T \cdot \mathbf{J}z_{n_\ell+1}\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\boldsymbol{\theta},\mathbf{z}) \\
&= (\nabla\phi(\langle [\![P_1]\!],\dots,[\![P_\ell]\!],\pi_{n_{\ell+1}}\rangle(\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\mathbf{z},\boldsymbol{\theta}))))^T \cdot \mathbf{e}_{n_\ell+1} \\
&= \frac{\partial \phi}{\partial z_{n_\ell+1}}(\langle [\![P_1]\!],\dots,[\![P_\ell]\!],\pi_{n_{\ell+1}}\rangle(\langle \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(1)},\dots \boldsymbol{\phi}_{\boldsymbol{\theta}}^{(n_\ell)},\pi_{n_{\ell+1}}\rangle(\mathbf{z},\boldsymbol{\theta}))) \\
&\neq 0
\end{aligned}$$

by assumption (and the infimum follows immediately, too).

The case for addition/multiplication uses the Weakening Lemma A.1 and the case for application uses the Substitution Lemma A.2. The case for conditionals uses Lemma A.3. □

Suppose $\emptyset \mid \emptyset \vdash_{\boldsymbol{\theta}} M : R$ is such that at most $n \in \mathbb{N}$ samples are drawn, i.e. $[\![M]\!] : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$. If $0, n' \rhd_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)} P$ then $n' \leq n$ and we also write $M \Downarrow_{\boldsymbol{\phi}'}^{(\Psi'_<,\Psi'_\geq)} P$, where the domains have been lifted from $\mathbb{R}^m \times \mathbb{R}^{n'}$ and $\mathbb{R}^{n'}$ to $\mathbb{R}^m \times \mathbb{R}^n$ and $\mathbb{R}^n$, respectively.

Proposition A.5 (Soundness and Completeness)**.** *If* $\emptyset \mid \emptyset \vdash_{\boldsymbol{\theta}} M : R$ *then*

$$[\![M]\!](\boldsymbol{\theta},\mathbf{z}) = \sum_{M \Downarrow_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)} P} [\![P]\!](\boldsymbol{\theta},\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \prod_{\psi\in\Psi_<} [\psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) < 0] \cdot \prod_{\psi\in\Psi_\geq} [\psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \geq 0]$$

The proof is very similar to (Mak et al., 2021, Thm. 1, Lem. 7). In our setting we do not have to deal with neither recursion nor the weight function, but we do have to take care of the transformations.

B. Supplementary Materials for Section 3

We focus on the following setting. The results of Section 3 will follow by linearity.

Assumption B.1.

1. $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is compact.
2. $p : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ and $\psi_1,\dots,\psi_\ell : \mathbb{R}^n \to \mathbb{R}$ are polynomials and the ψ_i are not constant 0.
3. For $j \in \{1,\dots,n\}$, $\phi_{\boldsymbol{\theta}}^{(j)} : \mathbb{R}^j \to \mathbb{R}^n$ is a diffeomorphic polynomial such that $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \inf_{z_1,\dots,z_j\in\mathbb{R}} \left|\frac{\partial \phi^{(j)}}{\partial z_j}(\boldsymbol{\theta},(z_1,\dots,z_j))\right| > 0$

Define

$$\begin{aligned}
\Psi(\boldsymbol{\alpha}) &\coloneqq \prod_{i=1}^{\ell'} [\psi_i(\boldsymbol{\alpha}) \geq 0] \cdot \prod_{i=\ell'+1}^{\ell} [\psi_i(\boldsymbol{\alpha}) > 0] & \Psi_k(\boldsymbol{\alpha}) &\coloneqq \prod_{i=1}^{\ell} \sigma_k(\psi_i(\boldsymbol{\alpha})) \\
\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) &\coloneqq \left\langle \phi_{\boldsymbol{\theta}}^{(1)}(z_1),\dots,\phi_{\boldsymbol{\theta}}^{(n)}(z_1,\dots,z_n) \right\rangle \\
f(\boldsymbol{\theta},\mathbf{z}) &\coloneqq p(\boldsymbol{\theta},\mathbf{z}) \cdot \Psi(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) & f_k(\boldsymbol{\theta},\mathbf{z}) &\coloneqq p(\boldsymbol{\theta},\mathbf{z}) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))
\end{aligned}$$

B.1. Supplementary Materials for Section 3.1

The following immediately follows from a well-known result about exchanging differentiation and integration, which is a consequence of the dominated convergence theorem (Klenke, 2014, Theorem 6.28):

Lemma B.2. *Let* $U \subseteq \mathbb{R}$ *be open. Suppose* $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ *satisfies*

1. *for each* $x \in \mathbb{R}$, $\mathbf{z} \mapsto g(x, \mathbf{z})$ *is integrable*
2. g *is continuously differentiable everywhere*
3. *there exists integrable* $h : \mathbb{R}^n \to \mathbb{R}$ *such that for all* $x \in U$ *and* $\mathbf{z} \in \mathbb{R}^n$, $|\frac{\partial g}{\partial x}(x, \mathbf{z})| \leq h(\mathbf{z})$.

Then for all $x \in U$, $\frac{\partial}{\partial x} \int g(x, \mathbf{z}) \, \mathrm{d}\mathbf{z} = \int \frac{\partial g}{\partial x}(x, \mathbf{z}) \, \mathrm{d}\mathbf{z}$.

Corollary B.3. *Let* $i \in \{1, \ldots, m\}$, $M > 0$ *and* $U \coloneqq B_M(\mathbf{0}) \subseteq \mathbb{R}^m$ *be the open* M*-ball. Suppose* $g : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ *satisfies*

1. *for each* $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{z} \mapsto g(\mathbf{x}, \mathbf{z})$ *is integrable*
2. g *is continuously differentiable everywhere*
3. *there exists integrable* $h : \mathbb{R}^n \to \mathbb{R}$ *such that for all* $\mathbf{x} \in U$ *and* $\mathbf{z} \in \mathbb{R}^n$, $|\frac{\partial g}{\partial x_i}(\mathbf{x}, \mathbf{z})| \leq h(\mathbf{z})$.

Then for all $\mathbf{x} \in U$, $\frac{\partial}{\partial x_i} \int g(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} = \int \frac{\partial g}{\partial x_i}(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z}$.

Proposition B.4. *For all* $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \mathbb{R}^m$ *and* $i \in \{1, \ldots, m\}$,

$$\frac{\partial}{\partial \theta_i} \mathbb{E}_{\mathbf{z}} \left[f_k(\boldsymbol{\theta}, \mathbf{z}) \right] = \mathbb{E}_{\mathbf{z}} \left[\frac{\partial}{\partial \theta_i} f_k(\boldsymbol{\theta}, \mathbf{z}) \right]$$

Proof. Suppose $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \mathbb{R}^m$ and $i \in \{1, \ldots, m\}$. Clearly, there exists $M > 0$ such that $\boldsymbol{\theta} \in B_M(\mathbf{0}) \subseteq \mathbb{R}^m$.

For $j \in \{1, \ldots, \ell\}$, abbreviate $p_j(\boldsymbol{\theta}, \mathbf{z}) \coloneqq \psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))$, which are polynomials. Note

$$\begin{aligned}
&\left| \frac{\partial}{\partial \theta_i} \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{j=1}^{\ell} \sigma_k(p_j(\boldsymbol{\theta}, \mathbf{z})) \right) \right| \\
&\leq \left| \frac{\partial p}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right| + \left| p(\boldsymbol{\theta}, \mathbf{z}) \cdot \sum_{j=1}^{\ell} \sigma_k'(p_j(\boldsymbol{\theta}, \mathbf{z})) \cdot \frac{\partial p_j}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{j' \neq j} \sigma(p_{j'}(\boldsymbol{\theta}, \mathbf{z})) \right| \\
&\leq \left| \frac{\partial p}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right| + \left| p(\boldsymbol{\theta}, \mathbf{z}) \cdot \sum_{j=1}^{\ell} \frac{\sqrt{k}}{4} \cdot \frac{\partial p_j}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right| \\
&\leq \left| \underbrace{\frac{\partial p}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z})}_{\text{polynomial}} \right| + \frac{\sqrt{k}}{4} \cdot \sum_{j=1}^{\ell} \left| \underbrace{p(\boldsymbol{\theta}, \mathbf{z}) \cdot \frac{\partial p_j}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z})}_{\text{polynomial}} \right|
\end{aligned}$$

because $\sigma_k' \leq \frac{\sqrt{k}}{4}$. By Lemma E.2 the summands can be uniformly bounded on the closure of $B_M(\mathbf{0})$ by the absolute value of an (integrable) polynomial. The claim follows by Corollary B.3. □

Proposition 3.1 follows by linearity.

B.2. Supplementary Materials for Section 3.2

Note that by Item 3 of Assumption B.1, $\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})$ is (lower) triangular and invertible. In particular,

Lemma B.5. *There exists* $M > 0$ *such that* $|\det \mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})| > M$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $\mathbf{z} \in \mathbb{R}^n$.

Hence,

$$\begin{aligned}
\boldsymbol{\chi}_i : \mathbb{R}^m \times \mathbb{R}^n &\to \mathbb{R}^n \\
(\boldsymbol{\theta}, \mathbf{z}) &\mapsto \mathbf{J}_{\mathbf{z}}^{-1} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \mathbf{J}_{\theta_i} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})
\end{aligned}$$

is well-defined.

As a consequence, we can express partial derivatives w.r.t. a parameter θ_i in terms of the gradient w.r.t. the latent variables $\mathbf{z}$.

$$\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) = \nabla_{\mathbf{z}}(\Psi_k \circ \boldsymbol{\phi}_{(-)})(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_i(\boldsymbol{\theta}, \mathbf{z}) \tag{5}$$

Note that for this it is essential that the guards, i.e. the ψ_i in the definition of Ψ_k do not depend on the parameters, but that all ψ_i are composed with the same transformation[6] $\boldsymbol{\phi}_{\boldsymbol{\theta}}$.

Lemma B.6. *Let $i, i' \in \{1, \dots m\}$ and $j, j' \in \{1, \dots, n\}$. All of $|\boldsymbol{\chi}_i|$, $|\frac{\partial \boldsymbol{\chi}_i}{\partial z_j}|$, $|\frac{\partial^2 \boldsymbol{\chi}_i}{\partial \theta_{i'} \partial z_j}|$ and $|\frac{\partial^2 \boldsymbol{\chi}_i}{\partial z_j \partial z_{j'}}|$ are bounded by the absolute value of polynomials.*

Proof. Let $M > 0$ as in Lemma B.5. The claim for $\boldsymbol{\chi}_i$ follows immediately from

$$\frac{1}{\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{z}))} \leq \frac{1}{M}$$

and the fact that adjugate (transpose of the cofactor matrix) is a polynomial. Similarly,

$$\left|\frac{\partial}{\partial z_j} \frac{1}{\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}\right| = \left|\frac{1}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^2} \cdot \frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right| \leq \frac{1}{M^2} \cdot \left|\underbrace{\frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{polynomial}}\right|$$

Similarly,

$$\begin{aligned}
&\left|\frac{\partial^2}{\partial z_{j'} \partial z_j} \frac{1}{\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}\right| \\
&= \left|\frac{\partial}{\partial z_{j'}} \left(\frac{1}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^2} \cdot \frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right)\right| \\
&\leq \left|\frac{1}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^2} \cdot \underbrace{\frac{\partial^2}{\partial z_{j'} \partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{polynomial}} - \frac{2}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^3} \cdot \underbrace{\frac{\partial}{\partial z_{j'}} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{polynomial}}\right|
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left|\frac{\partial^2}{\partial \theta_{i'} \partial z_j} \frac{1}{\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}\right| \\
&= \left|\frac{\partial}{\partial \theta_{i'}} \left(\frac{1}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^2} \cdot \frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right)\right| \\
&\leq \left|\frac{1}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^2} \cdot \underbrace{\frac{\partial^2}{\partial \theta_{i'} \partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{polynomial}} - \frac{2}{(\det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))^3} \cdot \underbrace{\frac{\partial}{\partial \theta_{i'}} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \frac{\partial}{\partial z_j} \det(\mathbf{J}_{\mathbf{z}} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{polynomial}}\right|
\end{aligned}$$

□

Lemma B.7. *Let $0 \neq q : \mathbb{R}^n \to \mathbb{R}$ be a polynomial. There exist $L, \epsilon > 0$ such that $\mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[|\sigma_{k+1}(q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - \sigma_k(q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))|^2] < L \cdot k^{-2-\epsilon}$ for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

[6] in this particular branch of the computation

Proof. Let $0 < \epsilon < \frac{1}{2}$.

Let $K, L, \delta > 0$ as in Corollary E.6. Suppose $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By Corollary E.6, there exists $U \subseteq \mathbb{R}^n$ such that $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\mathbf{z} \in \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U)]] \leq K \cdot k^{-\delta}$ and for each $\boldsymbol{\alpha} \in \mathbb{R}^n \setminus \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U)$, $|q(\boldsymbol{\alpha})| > L \cdot k^{-\epsilon}$. Thus,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left|\sigma_{k+1}(q_{\boldsymbol{\theta}}(\mathbf{z})) - \sigma_k(q_{\boldsymbol{\theta}}(\mathbf{z}))\right|^2\right] \\
&\leq \mathbb{E}_{\mathbf{z}}\left[[\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) \in U] \cdot \left|\sigma_{k+1}(q_{\boldsymbol{\theta}}(\mathbf{z})) - \sigma_k(q_{\boldsymbol{\theta}}(\mathbf{z}))\right|^2\right] + \mathbb{E}_{\mathbf{z}}\left[[\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathbb{R}^n \setminus U] \cdot \left|\sigma_{k+1}(q_{\boldsymbol{\theta}}(\mathbf{z})) - \sigma_k(q_{\boldsymbol{\theta}}(\mathbf{z}))\right|^2\right] \\
&\leq \mathbb{E}_{\mathbf{z}}\left[[\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) \in U] \cdot \frac{1}{k^2}\right] + \mathbb{E}_{\mathbf{z}}\left[[\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathbb{R}^n \setminus U] \cdot \left|\frac{1}{L} \cdot k^{-\frac{3}{2}+\epsilon}\right|^2\right] && \text{Lemmas F.1 and F.2} \\
&\leq \frac{1}{k^2} \cdot \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\mathbf{z} \in \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U)]] + \mathbb{E}_{\mathbf{z}}\left[\frac{1}{L^2} \cdot k^{-3+2\epsilon}\right] \\
&\leq k^{-2} \cdot K \cdot k^{-\delta} + L^{-2} \cdot k^{-3+2\epsilon} \\
&\leq \max\{K, L^{-2}\} \cdot k^{-2-\min\{\delta, 1-2\epsilon\}}
\end{aligned}
$$

□

As a consequence of Lemma B.7, Jensen's inequality, a telescoping series argument and the fact that the set of roots of polynomials is negligible (Caron, 2005) we obtain:

Corollary B.8. *Let $0 \neq q : \mathbb{R}^n \to \mathbb{R}$ be a polynomial. There exist $L, \epsilon > 0$ such that $\mathbb{E}_{\mathbf{z}}[|\sigma_k(q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - [q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) > 0]|] < L \cdot k^{-\epsilon}$ and $\mathbb{E}_{\mathbf{z}}[|\sigma_k(q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - [q(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \geq 0]|] < L \cdot k^{-\epsilon}$ for all k and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Note the following elementary fact that for $a_1, \ldots, a_\ell \in (0, 1)$ and $b_1, \ldots, b_\ell \in \{0, 1\}$,

$$
\left|\prod_{i=1}^{\ell} a_i - \prod_{i=1}^{\ell} b_i\right|^2 \leq \sum_{i=1}^{\ell} |a_i - b_i| \tag{6}
$$

Lemma B.9 (Uniform Convergence)**.** *Let $0 \neq q_1, \ldots, q_\ell : \mathbb{R}^n \to \mathbb{R}$ be polynomials and let the absolute value of $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be bounded by the absolute value of a polynomial.*

There exist $L, \epsilon > 0$ such that for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$
\mathbb{E}_{\mathbf{z}}[|h(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{i=1}^{\ell} \sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - h(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{i=1}^{\ell'} [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) > 0] \cdot \prod_{i=\ell'+1}^{\ell} [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \geq 0]|] < L \cdot k^{-\epsilon}
$$

Proof. By Corollary B.8 and Lemma E.3 there exist $\bowtie_i \in \{>, \geq\}$

$$
\mathbb{E}_{\mathbf{z}}[|h(\boldsymbol{\theta}, \mathbf{z})|^2] < L
$$

$$
\mathbb{E}_{\mathbf{z}}[|\sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \bowtie_i 0]|] < L \cdot k^{-\epsilon} \qquad \mathbb{E}_{\mathbf{z}}[|\sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \geq 0]|] < L \cdot k^{-\epsilon}
$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $i \in \{1, \ldots, \ell\}$. Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\left|h(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{i=1}^{\ell} \sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - h(\boldsymbol{\theta}, \mathbf{z}) \cdot \prod_{i=1}^{\ell} [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \bowtie_i 0]\right|\right] \\
&\leq \sqrt{\mathbb{E}_{\mathbf{z}}[|h(\boldsymbol{\theta}, \mathbf{z})|^2] \cdot \mathbb{E}_{\mathbf{z}}\left[\left|\prod_{i=1}^{\ell} \sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - \prod_{i=1}^{\ell} [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \bowtie_i 0]\right|^2\right]} && \text{Cauchy-Schwartz} \\
&\leq \sqrt{L \cdot \sum_{i=1}^{\ell} \mathbb{E}_{\mathbf{z}}\left[|\sigma_k(q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - [q_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \bowtie_i 0]|\right]} && \text{Equation (6)} \\
&\leq \sqrt{L \cdot \ell \cdot L \cdot k^{-\epsilon}} = L \cdot \sqrt{\ell} \cdot k^{-\frac{\epsilon}{2}}
\end{aligned}
$$

□

B.3. Proof of Proposition 3.3

Lemma B.10. *There exist $L, \epsilon > 0$ and integrable $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}_{\mathbf{z}}[|\frac{\partial f_k}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) - h(\boldsymbol{\theta}, \mathbf{z})|] < L \cdot k^{-\epsilon}$ for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Proof. First, note that uniform convergence of

$$E_{\mathbf{z}}\left[\underbrace{\left(\frac{\partial}{\partial \theta_i} p(\boldsymbol{\theta}, \mathbf{z})\right)}_{\text{polynomial}} \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right]$$

follows from Lemma B.9. Besides,

$$\begin{aligned}
&E_{\mathbf{z}}\left[p(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z})\right)\right] \\
&= \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z})\right)\right] && \text{Equation (5)} \\
&= \sum_{j=1}^{n} \int \mathcal{N}(\mathbf{z}_{-j}) \cdot \int \mathcal{N}(z_j) \cdot p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z})\right) \mathrm{d}z_j \, \mathrm{d}\mathbf{z}_{-j} \\
&= \sum_{j=1}^{n} \int \mathcal{N}(\mathbf{z}_{-j}) \cdot \underbrace{\left[\mathcal{N}(z_j) \cdot p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right]_{-\infty}^{\infty}}_{0} \mathrm{d}\mathbf{z}_{-j} && \text{integration by parts} \\
&\quad - \sum_{j=1}^{n} \int \mathcal{N}(\mathbf{z}_{-j}) \cdot \int \frac{\partial}{\partial z_j}\left(\mathcal{N}(z_j) \cdot p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \, \mathrm{d}z_j \, \mathrm{d}\mathbf{z}_{-j} \\
&= \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}[z_j \cdot p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))] && \frac{\partial \mathcal{N}}{\partial x}(x) = -x \cdot \mathcal{N}(x) \\
&\quad - \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right]
\end{aligned}$$

By Lemma B.6 $|\boldsymbol{\chi}_i|$ and $|\frac{\partial \boldsymbol{\chi}_i}{\partial z_j}|$ and are bounded by the absolute value of polynomials and clearly $0 \leq \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \leq 1$. Therefore, uniform convergence of $\mathbb{E}[p(\boldsymbol{\theta}, \mathbf{z}) \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\theta_i}(\boldsymbol{\theta}, \mathbf{z})]$ follows again from Lemma B.9. □

The following is an instance of (Rudin, 1973, Thm. 7.17):

Lemma B.11. *Suppose $g : [a, b] \to \mathbb{R}$ and $g_k : [a, b] \to \mathbb{R}$ (for $k \in \mathbb{N}$) satisfy*

1. *each g_k is differentiable*
2. *$g(x) = \lim_{k\to\infty} g_k(x)$ for all $x \in [a, b]$*
3. *$g'_k(x)$ converges uniformly on $[a, b]$.*

Then $g'(x) = \lim_{k\to\infty} g'_k(x)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Corollary B.12. *Let $U \subseteq \mathbb{R}^m$ be compact and $i \in \{1, \ldots, m\}$. Suppose $g : \mathbb{R}^m \to \mathbb{R}$ and $g_k : \mathbb{R}^m \to \mathbb{R}$ (for $k \in \mathbb{N}$) satisfy*

1. *each g_k is differentiable*

2. $g(\mathbf{x}) = \lim_{k\to\infty} g_k(\mathbf{x})$ *for all* $\mathbf{x} \in U$
3. $\frac{\partial g_k}{\partial x_i}$ *converges uniformly on* U.

Then $\frac{\partial g}{\partial x_i}(\mathbf{x}) = \lim_{k\to\infty} \frac{\partial g_k}{\partial x_i}(\mathbf{x})$ *for all* $\mathbf{x} \in U$.

Proposition B.13. *Let* $i \in \{1, \dots, m\}$. *There exist* $L, \epsilon > 0$ *such that for all* $k \in \mathbb{N}$ *and* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\left| \frac{\partial}{\partial \theta_i} \mathbb{E}_{\mathbf{z}}[f_k(\boldsymbol{\theta}, \mathbf{z})] - \frac{\partial}{\partial \theta_i} \mathbb{E}[f(\boldsymbol{\theta}, \mathbf{z})] \right| < L \cdot k^{-\epsilon}$$

Proof. Let $g_k(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{z}}[f_k(\boldsymbol{\theta}, \mathbf{z})]$ and $g(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{z}}[f(\boldsymbol{\theta}, \mathbf{z})]$. By Proposition 3.1 each g_k is differentiable. By Lemma B.9, $g(\boldsymbol{\theta}) = \lim_{k\to\infty} g_k(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By Lemma B.10, $\frac{\partial g_k}{\partial \theta_i}$ converges uniformly on $\boldsymbol{\Theta}$ to some $\mathbb{E}_{\mathbf{z}}[h(\boldsymbol{\theta}, \mathbf{z})]$. Therefore, by Corollary B.12,

$$\frac{\partial}{\partial \theta_i} \mathbb{E}_{\mathbf{z}}[f(\boldsymbol{\theta}, \mathbf{z})] = \frac{\partial g}{\partial \theta_i}(\boldsymbol{\theta}) = \lim_{k\to\infty} \frac{\partial g_k}{\partial \theta_i}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}}[h(\boldsymbol{\theta}, \mathbf{z})]$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore, by Lemma B.10,

$$\left| \frac{\partial}{\partial \theta_i} \mathbb{E}_{\mathbf{z}}[f_k(\boldsymbol{\theta}, \mathbf{z})] - \frac{\partial}{\partial \theta_i} \mathbb{E}[f(\boldsymbol{\theta}, \mathbf{z})] \right| = \left| \mathbb{E}_{\mathbf{z}} \left[\frac{\partial f_k}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right] - \mathbb{E}[h(\boldsymbol{\theta}, \mathbf{z})] \right| \leq \mathbb{E}_{\mathbf{z}} \left[\left| \frac{\partial f_k}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) - h(\boldsymbol{\theta}, \mathbf{z}) \right| \right] \leq L \cdot k^{-\epsilon}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. □

Proposition 3.3 follows.

C. Supplementary Materials for Section 4

Proposition 4.1 (Convergence)**.** *Suppose* $\gamma_k \in \Theta(1/k)$ *and* $g_k(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{z}}[f_k(\boldsymbol{\theta}, \mathbf{z})]$ *and* $g(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\mathbf{z}}[f(\boldsymbol{\theta}, \mathbf{z})]$ *are well-defined and differentiable.*

Suppose there exist $\{\boldsymbol{\theta}_k \mid k \in \mathbb{N}\} \subseteq \boldsymbol{\Theta} \subseteq \mathbb{R}^m$, $L > 0$ *and* $\epsilon > 0$ *s.t. for all* $k \in \mathbb{N}$ *and* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

(D1) $\nabla_{\boldsymbol{\theta}} g_k(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}}[\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{z})]$

(D2) $|g_{k+1}(\boldsymbol{\theta}) - g_k(\boldsymbol{\theta})| < k^{-1-\epsilon} \cdot L$

(D3) $\|\nabla g_k(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta})\|^2 < k^{-\epsilon} \cdot L$

(D4) $\mathbb{E}_{\mathbf{z}}[\|\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{z})\|^2] < L$

(D5) $\|\mathbf{H}\, g_k(\boldsymbol{\theta})\| < L$

Then $\inf_{i\in\mathbb{N}} \mathbb{E}[\|\nabla g(\boldsymbol{\theta}_i)\|^2] = 0$.

Proof. By Taylor's theorem,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}_k}[g_{k+1}(\boldsymbol{\theta}_{k+1})] \\
&\leq \mathbb{E}_{\mathbf{z}_k}[g_k(\boldsymbol{\theta}_{k+1})] + k^{-1-\epsilon} \cdot L && \text{(D2)} \\
&\leq \mathbb{E}_{\mathbf{z}_k}\left[g_k(\boldsymbol{\theta}_k) - \gamma_k \cdot \langle \nabla_{\boldsymbol{\theta}} g_k(\boldsymbol{\theta}_k), \nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}_k, \mathbf{z}_k) \rangle + \frac{\gamma_k^2}{2} \cdot \|\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}_k, \mathbf{z}_k)\|^2 \cdot \|\mathbf{H}\, g_k\| \right] + k^{-1-\epsilon} \cdot L && \text{Taylor's theorem} \\
&\leq g_k(\boldsymbol{\theta}_k) - \gamma_k \cdot \|\nabla_{\boldsymbol{\theta}} g_k(\boldsymbol{\theta}_k)\|^2 + \frac{\gamma_k^2}{2} \cdot L^2 + k^{-1-\epsilon} \cdot L && \text{(D1), (D4) and (D5)} \\
&\leq g_k(\boldsymbol{\theta}_k) - \gamma_k \cdot \|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_k)\|^2 + \gamma_k \cdot k^{-\epsilon} \cdot L + \frac{\gamma_k^2}{2} \cdot L^2 + k^{-1-\epsilon} \cdot L && \text{(D3)}
\end{aligned}$$

Hence,

$$\gamma_k \|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_k)\|^2 \leq g_k(\boldsymbol{\theta}_k) - \mathbb{E}_{\mathbf{z}_k}[g_{k+1}(\boldsymbol{\theta}_{k+1})] + r_k$$

where $r_k \coloneqq \gamma_k \cdot k^{-\epsilon} \cdot L + \frac{\gamma_k^2}{2} \cdot L^2 + k^{-1-\epsilon} \cdot L$. Hence,

$$\left(\sum_{i=0}^{k} \alpha_i\right) \cdot \min_{i=0}^{k} \mathbb{E}[\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_i)\|^2] \leq \sum_{i=0}^{k} \alpha_i \mathbb{E}[\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_i)\|^2] \leq g_0(\boldsymbol{\theta}_0) - \mathbb{E}[g_{k+1}(\boldsymbol{\theta}_{k+1})] + \sum_{i=0}^{k} r_i$$

where the expectation is taken w.r.t. $\mathbf{z}_0, \ldots, \mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that due to $\gamma_k \in \Theta(1/k)$, $\sum_{i\in\mathbb{N}} \alpha_i = \infty$ and $\sum_{i\in\mathbb{N}} r_i < \infty$. Besides, due to $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, which is bounded, the $g_i(\boldsymbol{\theta}_i)$ are bounded from below. Therefore, the claim follows immediately. □

C.1. Supplementary Materials for Section 4.1

Recall the framework set out at the beginning of Appendix B.

Proposition C.1 (Uniform Convergence)**.** *There exist $L, \epsilon > 0$ such that $\mathbb{E}_{\mathbf{z}}[|f_{k+1}(\boldsymbol{\theta}, \mathbf{z}) - f_k(\boldsymbol{\theta}, \mathbf{z})|] < L \cdot k^{-1-\epsilon}$ for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Proof. By Lemmas E.3 and B.7, there exist $L, \epsilon > 0$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $i \in \{1, \ldots, \ell\}$,

$$\mathbb{E}\left[|p(\boldsymbol{\theta}, \mathbf{z})|^2\right] < L \qquad \mathbb{E}_{\mathbf{z}}\left[\left|\sigma_{k+1}(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - \sigma_k(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right|^2\right] < L \cdot k^{-2-\epsilon}$$

Therefore, using the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}[|f_{k+1}(\boldsymbol{\theta}, \mathbf{z}) - f_k(\boldsymbol{\theta}, \mathbf{z})|] \\
&\leq \sum_{i=1}^{\ell} \mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\prod_{j=i}^{\ell} \sigma_{k+1}(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right) \cdot \left(\prod_{j=1}^{i-1} \sigma_k(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right) - p(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\prod_{j=i+1}^{\ell} \sigma_{k+1}(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right) \cdot \left(\prod_{j=1}^{i} \sigma_k(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right)\right|\right] \\
&\leq \sum_{i=1}^{\ell} \mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(\prod_{j=i+1}^{\ell} \sigma_{k+1}(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right) \cdot \left(\prod_{j=1}^{i-1} \sigma_k(\psi_j(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right) \cdot \left(\sigma_{k+1}(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - \sigma_k(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right)\right|\right] \\
&\leq \sum_{i=1}^{\ell} \sqrt{\mathbb{E}_{\mathbf{z}}\left[|p(\boldsymbol{\theta}, \mathbf{z})|^2\right] \cdot \mathbb{E}_{\mathbf{z}}\left[\left|\sigma_{k+1}(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))) - \sigma_k(\psi_i(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})))\right|^2\right]} \\
&\leq \ell \cdot \sqrt{L \cdot L \cdot k^{-2-\epsilon}} \leq \ell \cdot L \cdot k^{-1-\frac{\epsilon}{2}}
\end{aligned}$$

□

Proposition C.2 (Uniform Bound on Hessian)**.** *Let $i, i' \in \{1, \ldots, m\}$. There exists $L > 0$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial^2 f_k}{\partial\theta_{i'}\,\partial\theta_i}(\boldsymbol{\theta}, \mathbf{z})\right|\right] < L$$

Proof. First, note that

$$\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial^2 p}{\partial\theta_{i'}\,\partial\theta_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right|\right] \leq \mathbb{E}_{\mathbf{z}}\left[\left|\underbrace{\frac{\partial^2 p}{\partial\theta_{i'}\,\partial\theta_i}(\boldsymbol{\theta}, \mathbf{z})}_{\text{polynomial}}\right|\right]$$

and the right-hand side can be bounded uniformly by Lemma E.3. Besides,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\leq\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right|\right] && \text{Equation (5)}\\
&\leq\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|\left(z_j\cdot\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{z}))\right|\right] && \text{integration by parts}\\
&\leq\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|z_j\cdot\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(\frac{\partial p}{\partial \theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right|\right]
\end{aligned}$$

and the expectations on right-hand side can be bounded uniformly by Lemmas E.3 and B.6. (To be completely rigorous, to justify the second step we note that $z_j\mapsto\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})$ is a smooth function with a finite number of roots and invoke Lemma G.3.)

Finally, applying Equation (5) in the first and third step,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial^2\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial\theta_{i'}\,\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\leq\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial}{\partial\theta_{i'}}\left(\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right|\right]\\
&\leq\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial^2\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_j\,\partial\theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right|\right]+\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\boldsymbol{\chi}_{i,j}}{\partial\theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\leq\sum_{j=1}^{n}\sum_{j'=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial}{\partial z_j}\left(\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_{j'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right)\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\quad+\sum_{j=1}^{n}\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial\boldsymbol{\chi}_{i,j}}{\partial\theta_{i'}}(\boldsymbol{\theta},\mathbf{z})\right|\right] \qquad (7)
\end{aligned}$$

To bound the terms in the first sum uniformly, note that for each $j,j'\in\{1,\dots,n\}$, using integration by parts twice,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta},\mathbf{z})\cdot\frac{\partial}{\partial z_j}\left(\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_{j'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right)\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\leq\mathbb{E}_{\mathbf{z}}\left[\left|\left(z_j\cdot p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\frac{\partial\left(\Psi_k\circ\boldsymbol{\phi}_{(-)}\right)}{\partial z_{j'}}(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\leq\mathbb{E}_{\mathbf{z}}\left[\left|z_{j'}\cdot\left(z_j\cdot p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\quad+\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial}{\partial z_{j'}}\left(\left(z_j\cdot p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right)\cdot\Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right|\right]\\
&\leq\mathbb{E}_{\mathbf{z}}\left[\left|z_{j'}\cdot\left(z_j\cdot p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right|\right]\\
&\quad+\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial}{\partial z_{j'}}\left(\left(z_j\cdot p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})+\frac{\partial}{\partial z_j}\left(p(\boldsymbol{\theta},\mathbf{z})\cdot\boldsymbol{\chi}_{i,j}(\boldsymbol{\theta},\mathbf{z})\right)\right)\cdot\boldsymbol{\chi}_{i',j'}(\boldsymbol{\theta},\mathbf{z})\right)\right|\right]
\end{aligned}$$

and the expectations can be bounded uniformly by Lemmas E.3 and B.6.

The other expectations in Equation (7) can be bounded similarly. □

Proposition C.3 (Uniform Bound on "Variance"). *Let $i \in \{1, \dots, m\}$. There exists $L > 0$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial f_k}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z})\right|^2\right] < L$$

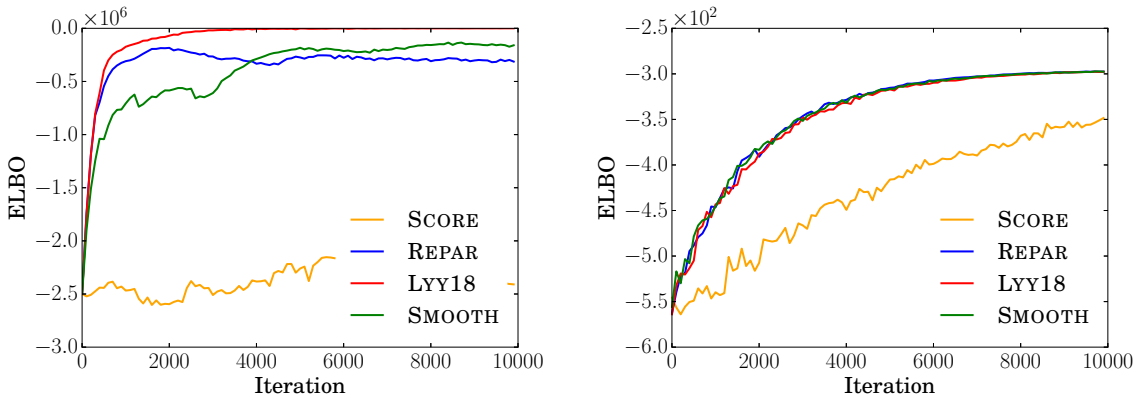
Proof.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta}, \mathbf{z}) \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z})\right|^2\right] \\
&\leq n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|p(\boldsymbol{\theta}, \mathbf{z}) \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right|^2\right] && \text{Equation (5)} \\
&\leq n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|z_j \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2 \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right|\right] \\
&\quad + n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial}{\partial z_j}\left(\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right|\right] && \text{integration by parts} \\
&\leq n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|z_j \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right|\right] \\
&\quad + n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial}{\partial z_j}\left(\frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right)\right|\right] \\
&\leq n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|\left(z_j^2 \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2 + \frac{\partial}{\partial z_j}\left(z_j \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right)\right) \cdot \Psi_k(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z}))\right|\right] \\
&\quad + n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|z_j \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right|\right] && \text{integration by parts} \\
&\leq n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|z_j^2 \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right|\right] + n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|\frac{\partial}{\partial z_j}\left(z_j \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right)\right|\right] \\
&\quad + n \cdot \sum_{j=1}^{n} \mathbb{E}_{\mathbf{z}}\left[\left|z_j \cdot \frac{\partial\left(\Psi_k \circ \boldsymbol{\phi}_{(-)}\right)}{\partial z_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot \left(p(\boldsymbol{\theta}, \mathbf{z}) \cdot \boldsymbol{\chi}_{i,j}(\boldsymbol{\theta}, \mathbf{z})\right)^2\right|\right]
\end{aligned}$$

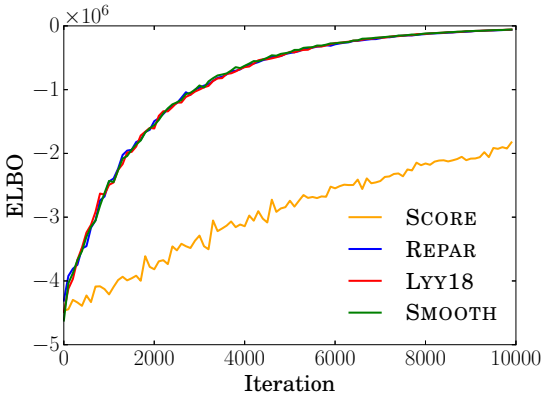
The terms in the third sum can be bounded uniformly as before by once again using integration by parts. For the others it follows directly from Lemmas E.3 and B.6. □

D. Supplementary Materials for Section 6

Results for publicly available code from (Lee et al., 2018) Building on the Python implementation of black-box variational inference given in Lee et al. (2018), we implement our smoothed gradient estimator and re-run their experiments with this additional estimator.



(a) `temperature` (b) `textmsg`



(c) `influenza`

Figure 5. ELBO trajectories for each model using the code from Lee et al. (2018) for $N = 1$.

The ELBO trajectories are shown in Figure 5 and the work-normalised variance ratios are given in Table 2 as the same way as in Section 6.

Some discrepancies between our implementation and this likely arise from how the models are implemented, ours written directly in Python and efficiently vectorised whereas the ones from Lee et al. (2018) are dynamically constructed by parsing a variant of lambda calculus with little to no optimisation.

The smoothed estimator is also not implemented quite as efficiently here since there is more overhead for constructing a gradient that explores all of the branches in their implementation.

Despite this, our smoothed estimator still performs comparably on the ELBO trajectories and has a smaller work-normalised variance in most situations.

Accuracy coefficients for the SMOOTH estimator For the SMOOTH estimator, for iteration $k \in \mathbb{N}$ we use the slightly modified $\sigma_k(x) \coloneqq \sigma(10\sqrt{k} \cdot x)$. The coefficient (here 10) was chosen to work well in the experimental setup given. Larger values resulted in a lower variance for each estimate but ultimately performed poorer when doing gradient descent with Adam. Intuitively, this comes from Adam being momentum-based and the a large smoothing value increases the bias of the initial few gradient estimates which are done on smoother approximations of the target function.

The experimental results for this on the `temperature` model are shown in Figure 6.

Table 2. Computational cost and work-normalised variances using the results from the Lee et al. (2018) code, all given as ratios with respect to the SCORE estimator (omitted since it would be all 1s).

(a) Temperature

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	1.12e+00	5.29e-09	2.81e-08
LYY18	1.23e+00	1.45e-08	5.12e-07
SMOOTH	1.08e+00	1.48e-09	1.03e-08

(b) Textmsg

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	1.15e+00	2.41e-02	2.44e-02
LYY18	1.19e+00	3.16e-02	2.97e-02
SMOOTH	1.11e+00	2.19e-02	2.16e-02

(c) Influenza

Estimator	Cost	Avg$(V(.))$	$V(\|\|.\|\|_2)$
REPARAM	2.27e+00	1.03e-02	4.91e-03
LYY18	2.41e+00	1.23e-02	5.83e-03
SMOOTH	3.63e+00	1.61e-02	7.77e-03

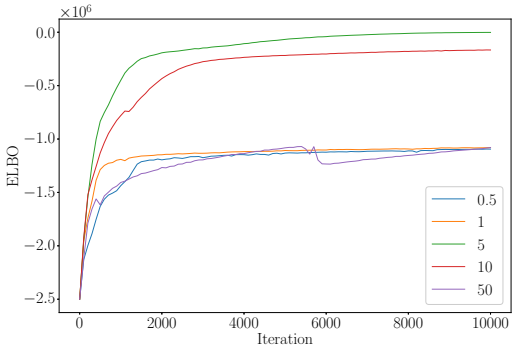


Figure 6. Varying the smoothing coefficient on the `temperature` model.

E. Properties of Polynomials

Assumption E.1. Throughout this section we assume that $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is compact.

E.1. Uniform Bounds

Lemma E.2. *For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $p_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}$ be a polynomial the coefficients of which depend continuously on $\boldsymbol{\theta}$. Then there exists a polynomial $q : \mathbb{R}^n \to \mathbb{R}$ such that $|p_{\boldsymbol{\theta}}| \leq |q|$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Proof. First, suppose $f : \boldsymbol{\Theta} \to \mathbb{R}$, the coefficient function of a monomial, is continuous and let $k_1, \ldots, k_n \in \mathbb{N}$. Since, $|f|$ is continuous and $\boldsymbol{\Theta}$ is compact there exists $M := \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |f(\boldsymbol{\theta})|$. Then

$$\left| f(\boldsymbol{\theta}) \cdot \prod_{i=1}^{n} z_i^{k_i} \right| \leq M \cdot |z_i^{k_i}|$$

The claim follows by linearity. □

Lemma E.3. *For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $p_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}$ be a polynomial the coefficients of which depend continuously on $\boldsymbol{\theta}$. Then there exists $M > 0$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[|p_{\boldsymbol{\theta}}(\mathbf{z})|] < M$.*

Proof. By Lemma E.2 there exists a polynomial $q : \mathbb{R}^n \to \mathbb{R}$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $|p_{\boldsymbol{\theta}}| \leq |q|$. Furthermore, (absolute) moments of normals are finite. Therefore $|q|$ is integrable and $M := \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[|q(\mathbf{z})|]$ is well-defined. □

E.2. Behaviour Near Roots

Lemma E.4. *Let $\epsilon > 0$ and $p(z) = \sum_{i=0}^{n} a_i \cdot z^i$ be a (univariate) polynomial such that $a_n \neq 0$.*

Then for all $k \in \mathbb{N}$ there exist $U \subseteq \mathbb{R}$ such that $\mathrm{Leb}(U) \leq 2n \cdot k^{-\frac{\epsilon}{n}}$ and for all $z \in \mathbb{R} \setminus U$, $|p(z)| > a_n \cdot k^{-\epsilon}$.

Proof. By the fundamental theorem of algebra p has the form $p(z) = a_n \cdot \prod_{i=1}^{n}(z - b_i)$, where each $b_i \in \mathbb{C}$.

Now, suppose $k \in \mathbb{N}$. Define $U \coloneqq \bigcup_{i=1}^{n}\{z \in \mathbb{R} \mid |z - b_i| \leq k^{-\frac{\epsilon}{n}}\}$. Clearly, $\mathrm{Leb}(U) \leq 2n \cdot k^{-\frac{\epsilon}{n}}$. Furthermore, for $z \in \mathbb{R} \setminus U$,

$$|p(z)| = |a_n| \cdot \prod_{i=1}^{n} |z - b_i| > L \cdot \prod_{i=1}^{n} k^{-\frac{\epsilon}{n}} = L \cdot k^{-\epsilon} \qquad \square$$

Lemma E.5. *Let $M > 0$ and for $j \in \{1, \ldots, n\}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $\nu_{\boldsymbol{\theta}}^{(j)}$ be a probability measure on $\mathbb{R}^j$ satisfying*

1. *for measurable $U \subseteq \mathbb{R}^j$, $\nu_{\boldsymbol{\theta}}^{(j+1)}(U \times \mathbb{R}) = \nu_{\boldsymbol{\theta}}^{(j)}(U)$*
2. *if for each $x_1, \ldots, x_j \in \mathbb{R}$ and $U_{(x_1,\ldots,x_j)} \subseteq \mathbb{R}$ is such that $\mathrm{Leb}(U_{(x_1,\ldots,x_j)}) \leq N$ then $\nu_{\boldsymbol{\theta}}^{(j+1)}(\{(x_1, \ldots, x_{j+1}) \mid x_{j+1} \in U_{(x_1,\ldots,x_j)}\}) \leq M \cdot N$.*

Furthermore, let $0 \neq p : \mathbb{R}^n \to \mathbb{R}$ be a polynomial and $\epsilon > 0$.

Then there exists $K, L, \delta > 0$ such that for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ there exist $U \subseteq \mathbb{R}^n$ such that $\nu_{\boldsymbol{\theta}}^{(n)}(U) \leq K \cdot k^{-\delta}$ and for all $\mathbf{x} \in \mathbb{R}^n \setminus U$, $|p(\mathbf{x})| > L \cdot k^{-\epsilon}$.

Proof. We prove the claim by induction on n. For $n = 0$ this is trivial. Hence, suppose $n \geq 0$, $\epsilon > 0$ and

$$p(x_1, \ldots, x_{n+1}) = \sum_{i=0}^{\ell} p_i(x_1, \ldots, x_n) \cdot x_{n+1}^i = p_{(x_1,\ldots,x_n)}(x_{n+1})$$

where each $p_i : \mathbb{R}^n \to \mathbb{R}$ is a polynomial and $p_\ell \neq 0$.

Let $K_\ell, L_\ell, \delta_\ell > 0$ be the constants the inductive hypothesis yields for p_ℓ. Define

$$K \coloneqq M \cdot K_\ell + 2\ell \cdot M \qquad L_\ell \coloneqq L \qquad \delta \coloneqq \min\left\{\delta_\ell, \frac{\epsilon}{2\ell}\right\}$$

Suppose $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For each $(x_1, \ldots, x_n) \in \mathbb{R}^n$, applying Lemma E.4 to $p_{(x_1,\ldots,x_n)}$ and $\frac{\epsilon}{2}$ there exists $U_{(x_1,\ldots,x_n)} \subseteq \mathbb{R}$ such that

$$\mathrm{Leb}\left(U_{(x_1,\ldots,x_n)}\right) \leq 2\ell \cdot k^{-\frac{\epsilon}{2\ell}} \tag{8}$$

$$p_{(x_1,\ldots,x_n)}(x_{n+1}) > |p_\ell(x_1, \ldots, x_n)| \cdot k^{-\frac{\epsilon}{2}} \tag{9}$$

for all $x_{n+1} \in \mathbb{R} \setminus U_{(x_1,\ldots,x_n)}$.

The inductive hypothesis for p_ℓ and $\frac{\epsilon}{2}$ yields $U_\ell \subseteq \mathbb{R}^n$ such that $\nu_{\boldsymbol{\theta}}^{(n)}(U_\ell) \leq K_\ell \cdot k^{-\delta}$ and for all $(x_1, \ldots, x_n) \in \mathbb{R}^n \setminus U$, $|p_\ell(x_1, \ldots, x_n)| > L_\ell \cdot k^{-\frac{\epsilon}{2}}$.

Define

$$U \coloneqq \left\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid (x_1, \ldots, x_n) \in U_\ell \text{ or } x_{n+1} \in U_{(x_1,\ldots,x_n)}\right\}$$

Note

$$\begin{aligned}
\nu_{\boldsymbol{\theta}}^{(n+1)}(U) &\leq \nu_{\boldsymbol{\theta}}^{(n+1)}(U_\ell \times \mathbb{R}) + \nu_{\boldsymbol{\theta}}^{(n+1)}(\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_{n+1} \in U_{(x_1,\ldots,x_n)}\}) \\
&\leq M \cdot K_\ell \cdot k^{-\delta_\ell} + 2M \cdot \ell \cdot k^{-\frac{\epsilon}{2\ell}} \\
&\leq K \cdot k^{-\delta}
\end{aligned}$$

Besides, if $(x, \dots, x_{n+1}) \in \mathbb{R}^{n+1} \setminus U$ then $(x_1, \dots, x_n) \in \mathbb{R}^n \setminus U_\ell$ and $x_{n+1} \in \mathbb{R} \setminus U_{(x_1,\dots,x_n)}$. Therefore,

$$|p(x_1, \dots, x_{n+1})| = |p_{(x_1,\dots,x_n)}(x_{n+1})| > |p_\ell(x_1, \dots, x_n)| \cdot k^{-\frac{\epsilon}{2}} > L_\ell \cdot k^\epsilon = L \cdot k^{-\epsilon} \qquad \square$$

Corollary E.6. *Let $\phi_{\boldsymbol{\theta}}^{(1)}, \dots, \phi_{\boldsymbol{\theta}}^{(n)}$ be as in Item 3 of Assumption B.1, let $0 \neq p : \mathbb{R}^n \to \mathbb{R}$ be a polynomial and $\epsilon > 0$. Then there exists $K, L, \delta > 0$ such that for all $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ there exist $U \subseteq \mathbb{R}^n$ such that $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\mathbf{z} \in \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U)]] \leq K \cdot k^{-\delta}$ and for all $\mathbf{x} \in \mathbb{R}^n \setminus U$, $|p(\mathbf{x})| > L \cdot k^{-\epsilon}$.*

Proof. For $j \in \{1, \dots, n\}$, let $\nu^{(j)}$ to be the probability measure on $\mathbb{R}^j$ with density

$$(z_1, \dots, z_j) \mapsto \prod_{i=1}^{j} \mathcal{N}\left(\left(\phi_{\boldsymbol{\theta}}^{(i)}\right)^{-1}(z_1, \dots, z_i)\right) \cdot \frac{\partial \phi^{(i)}}{\partial z_i}(\boldsymbol{\theta}, (z_1, \dots, z_i))$$

Note that $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[[\mathbf{z} \in \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(U)]] = \nu_{\boldsymbol{\theta}}(U)$ and by Lemma B.5 the conditions of Lemma E.5 are satisfied. $\square$

F. Properties of Sigmoid

Note that for $k \in \mathbb{N}$,

$$\sqrt{k+1} - \sqrt{k} \leq \frac{1}{\sqrt{k}} \tag{10}$$

Besides, for $x \in \mathbb{R}$ and $y \in \mathbb{R}_+$,

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \leq 1 - \sigma(x) \tag{11}$$

$$1 - \sigma(y) \leq \frac{1}{y} \tag{12}$$

$$1 - \sigma(y) \leq \frac{1}{y^2} \tag{13}$$

Lemma F.1. *Let $k \in \mathbb{N}$ and $y \in \mathbb{R}$, then $|\sigma_{k+1}(y) - \sigma_k(y)| \leq \frac{1}{k}$.*

Proof. Note that $|\sigma_{k+1}(-y) - \sigma_k(-y)| = |1 - \sigma_{k+1}(y) - (1 - \sigma_k(y))|$. Therefore, it suffices to consider $y > 0$. By Taylor's theorem there exists $\zeta \in (\sqrt{k} \cdot y, \sqrt{k+1} \cdot y)$ such that

$$\begin{aligned}
|\sigma_{k+1}(y) - \sigma_k(y)| &= \sigma'(\zeta) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y \\
&\leq \sigma'(\sqrt{k} \cdot y) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \sigma'_{|\mathbb{R}_+} \text{ decreasing} \\
&\leq (1 - \sigma(\sqrt{k} \cdot y)) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \text{Equation (11)} \\
&\leq \frac{1}{\sqrt{k} \cdot y} \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \text{Equation (12)} \\
&\leq \frac{1}{\sqrt{k} \cdot y} \cdot \frac{1}{\sqrt{k}} \cdot y && \text{Equation (10)} \\
&= \frac{1}{k}
\end{aligned}$$

$\square$

Lemma F.2. *Let $L, \epsilon > 0$, $k \in \mathbb{N}$ and $y \in \mathbb{R}$ be such that $|y| > L \cdot k^{-\epsilon}$, then $|\sigma_{k+1}(y) - \sigma_k(y)| \leq \frac{1}{L} \cdot k^{-\frac{3}{2}+\epsilon}$.*

Proof. Again, it suffices to consider $y > 0$. By Taylor's theorem there exists $\zeta \in (\sqrt{k} \cdot y, \sqrt{k+1} \cdot y)$ such that

$$\begin{aligned}
|\sigma_{k+1}(y) - \sigma_k(y)| &= \sigma'(\zeta) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y \\
&\leq \sigma'(\sqrt{k} \cdot y) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \sigma'_{|\mathbb{R}_+} \text{ decreasing} \\
&\leq (1 - \sigma(\sqrt{k} \cdot y)) \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \text{Equation (11)} \\
&\leq \frac{1}{\left(\sqrt{k} \cdot y\right)^2} \cdot (\sqrt{k+1} - \sqrt{k}) \cdot y && \text{Equation (13)} \\
&\leq \frac{1}{\left(\sqrt{k} \cdot y\right)^2} \cdot \frac{1}{\sqrt{k}} \cdot y && \text{Equation (10)} \\
&\leq \frac{1}{k \cdot \sqrt{k} \cdot y} \\
&\leq \frac{1}{k \cdot \sqrt{k} \cdot L \cdot k^{-\epsilon}} && L \cdot k^{-\epsilon} < y
\end{aligned}$$

□

G. Integration by Parts

Lemma G.1. *Suppose $f, g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are smooth. Then for every $\theta \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z) \, \mathrm{d}z = [f(\theta, z) \cdot g(\theta, z)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) \, \mathrm{d}z$$

Proof. By the product rule, for every $\theta, z \in \mathbb{R}$,

$$\frac{\partial}{\partial z} \left(f(\theta, z) \cdot g(\theta, z) \right) = \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) + f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z)$$

Therefore, for every $\eta > 0$,

$$\int_{-\eta}^{\eta} \frac{\partial}{\partial z} \left(f(\theta, z) \cdot g(\theta, z) \right) \mathrm{d}z = \int_{-\eta}^{\eta} \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) \, \mathrm{d}z + \int_{-\eta}^{\eta} f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z) \, \mathrm{d}z$$

Besides, for every $\eta > 0$,

$$\int_{-\eta}^{\eta} \frac{\partial}{\partial z} \left(f(\theta, z) \cdot g(\theta, z) \right) \mathrm{d}z = [f(\theta, z) \cdot g(\theta, z)]_{-\eta}^{\eta}$$

Therefore, provided the limits exist,

$$\begin{aligned}
[f(\theta, z) \cdot g(\theta, z)]_{-\infty}^{\infty} &= \lim_{\eta \to \infty} \int_{-\eta}^{\eta} \frac{\partial}{\partial z} \left(f(\theta, z) \cdot g(\theta, z) \right) \mathrm{d}z \\
&= \lim_{\eta \to \infty} \left(\int_{-\eta}^{\eta} \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) \, \mathrm{d}z + \int_{-\eta}^{\eta} f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z) \, \mathrm{d}z \right) \\
&= \int_{-\infty}^{\infty} \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) \, \mathrm{d}z + \int_{-\infty}^{\infty} f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z) \, \mathrm{d}z
\end{aligned}$$

and the claim follows. □

Corollary G.2. *Suppose $f, g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are smooth and that for all $\theta \in \mathbb{R}$, $\lim_{z \to \pm\infty} (f(\theta, z) \cdot g(\theta, z)) = 0$. Then for every $\theta \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} f(\theta, z) \cdot \frac{\partial g}{\partial z}(\theta, z) \, \mathrm{d}z = - \int_{-\infty}^{\infty} \frac{\partial f}{\partial z}(\theta, z) \cdot g(\theta, z) \, \mathrm{d}z$$

Lemma G.3. *Let $M > 0$ and $f, g : \mathbb{R} \to \mathbb{R}$ be smooth such that $0 \leq f, f' \leq M$, $\lim_{x\to\pm\infty} g(x) = 0$ and there is a finite set of disjoint possibly unbounded open intervals $(a_1, b_1), \dots, (a_m, b_m)$ of $\mathbb{R}$ such that $g(x) < 0$ for $x \in (a_i, b_i)$ and $g(x) \geq 0$ for $x \notin \bigcup_{i=1}^n (a_i, b_i)$.*

Then $\int |f'(x) \cdot g(x)| \,\mathrm{d}x \leq \int |f(x) \cdot g'(x)| \,\mathrm{d}x$.

Proof. Abusing notation, set $b_0 \coloneqq a_{n+1} \coloneqq \infty$. Note that[7] $\mathbb{R} = \bigcup_{i=1}^n (a_i, b_i) \cup \bigcup_{i=0}^n [b_i, a_{i+1}]$. Besides, by continuity and assumption, $f(x) \cdot g(x) = 0$ for $x = a_i$ or $x = b_i$ (taking limits for $\pm\infty$). Therefore, by partial integration,

$$\begin{aligned}
\int_{\mathbb{R}} |f'(x) \cdot g(x)| \,\mathrm{d}x &= -\sum_{i=1}^n \int_{a_i}^{b_i} f'(x) \cdot g(x) \,\mathrm{d}x + \sum_{i=0}^n \int_{b_i}^{a_{i+1}} f'(x) \cdot g(x) \,\mathrm{d}x \\
&= \sum_{i=1}^n \left(\int_{a_i}^{b_i} f(x) \cdot g'(x) \,\mathrm{d}x - \underbrace{[f(x) \cdot g(x)]_{a_i}^{b_i}}_{0} \right) + \sum_{i=0}^n \left(\underbrace{[f(x) \cdot g(x)]_{b_i}^{a_{i+1}}}_{0} - \int_{b_i}^{a_{i+1}} f(x) \cdot g'(x) \,\mathrm{d}x \right) \\
&\leq \sum_{i=1}^n \int_{a_i}^{b_i} |f(x) \cdot g'(x)| \,\mathrm{d}x + \sum_{i=0}^n \int_{b_i}^{a_{i+1}} |f(x) \cdot g'(x)| \,\mathrm{d}x = \int_{\mathbb{R}} |f(x) \cdot g'(x)| \,\mathrm{d}x
\end{aligned}$$

□

[7] (Ab)using $[-\infty, a_1] \coloneqq (-\infty, b_1]$ and $[b_n, \infty] \coloneqq [b_n, \infty)$.