Diagonalisation SGD: Fast & Convergent SGD for Non-Differentiable Models via Reparameterisation and Smoothing

Dominik Wagner
NTU Singapore

Basim Khajwal
Jane Street

C.-H. Luke Ong
NTU Singapore

Abstract

It is well-known that the reparameterisation gradient estimator, which exhibits low variance in practice, is biased for non-differentiable models. This may compromise correctness of gradient-based optimisation methods such as stochastic gradient descent (SGD). We introduce a simple syntactic framework to define non-differentiable functions piecewisely and present a systematic approach to obtain smoothings for which the reparameterisation gradient estimator is unbiased. Our main contribution is a novel variant of SGD, *Diagonalisation Stochastic Gradient Descent*, which progressively enhances the accuracy of the smoothed approximation *during* optimisation, and we prove convergence to stationary points of the *unsmoothed* (original) objective. Our empirical evaluation reveals benefits over the state of the art: our approach is simple, fast, stable and attains orders of magnitude reduction in work-normalised variance.

1 INTRODUCTION

In this paper we investigate stochastic optimisation problems of the form

$$\operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[f_{\boldsymbol{\theta}}(\mathbf{z})] \tag{1}$$

which have a wide array of applications, ranging from variational inference and reinforcement learning to queuing theory and portfolio design (Mohamed et al., 2020; Blei et al., 2017; Zhang et al., 2019; Sutton and Barto, 2018).

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

We are interested in scenarios in which $f_{\boldsymbol{\theta}}$ is directly expressed in a programming language. Owing to the presence of if-statements, which arise naturally when modelling real-world problems (see Section 6), $f_{\boldsymbol{\theta}}$ may not be continuous, let alone differentiable.

In variational inference, Bayesian inference is framed as an optimisation problem and $f_{\boldsymbol{\theta}}$ is the *evidence lower bound* (ELBO) $\log p(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\theta}}(\mathbf{z})$, where p is the *model* and $q_{\boldsymbol{\theta}}$ is a *variational approximation*. Our prime motivation is the advancement of variational inference for probabilistic programming[1]—a new programming paradigm to pose and automatically solve Bayesian inference problems.

Gradient Based Optimisation. In practice the standard method to solve optimisation problems of the form (1) are variants of *Stochastic Gradient Descent (SGD)*. Since the objective function is in general not convex, we cannot hope to always find global optima and we seek *stationary points* instead, where the gradient w.r.t. the parameters $\boldsymbol{\theta}$ vanishes (Robbins and Monro, 1951).

A crucial ingredient for *fast* convergence to a *correct* stationary point is an estimator of gradients of the objective function $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[f_{\boldsymbol{\theta}}(\mathbf{z})]$ which is both unbiased and has low variance.

The *Score* or *REINFORCE* estimator (Ranganath et al., 2014; Wingate and Weber, 2013; Minh and Gregor, 2014) makes little assumptions about $f_{\boldsymbol{\theta}}$ but it frequently suffers from high variance resulting in suboptimal results or slow/unstable convergence.

An alternative approach is the *reparameterisation* or *pathwise* gradient estimator. The idea is to reparameterise the latent variables $\mathbf{z}$ in terms of a known base distribution (entropy source) via a diffeomorphic transformation $\boldsymbol{\phi}_{\boldsymbol{\theta}}$ (such as a location-scale transformation or cumulative distribution function). E.g. if

[1] see e.g. (van de Meent et al., 2018; Barthe et al., 2020) for introductions

$\mathcal{D}_{\boldsymbol{\theta}}(z)$ is a Gaussian $\mathcal{N}(z \mid \mu, \sigma^2)$ with $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ then the location-scale transformation using the standard normal as the base gives rise to the reparameterisation

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

In general, $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[f_{\boldsymbol{\theta}}(\mathbf{z})] = \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[f_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$ and its gradient can be estimated by (Mohamed et al., 2020):

$$\nabla_{\boldsymbol{\theta}}\, f_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})) \qquad \mathbf{s} \sim \mathcal{D}$$

It is folklore that the reparameterisation estimator typically exhibits significantly lower variance in practice than the score estimator (e.g. Mohamed et al. (2020); Rezende et al. (2014); Fu (2006); Schulman et al. (2015); Xu et al. (2019); Lee et al. (2018)). The reasons for this phenomenon are still poorly understood and examples do exist where score has lower variance than the reparameterisation estimator (Mohamed et al., 2020).

Unfortunately, the reparameterisation gradient estimator is biased for non-differentiable models, which can be easily expressed in a programming language by if-statements:

Example 1.1. We recall a simple counterexample (Lee et al., 2018, Prop. 2).

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

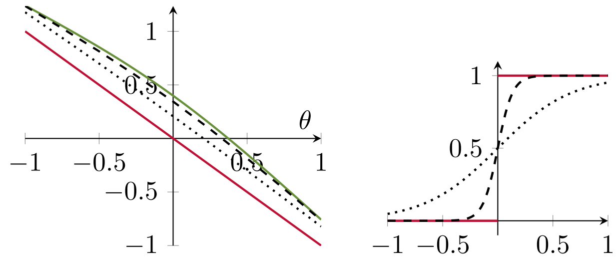
Observe that (see Fig. 1a):

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

Employing a biased gradient estimator may *compromise the correctness of stochastic optimisation*: even if we can find a point where the gradient estimator vanishes, it may not be a critical point of the objective function (1). Consequently, in practice we may obtain noticeably inferior results (see our experiments later, Fig. 2).

Systematic Smoothing. Khajwal et al. (2023) present a smoothing approach to avoid the bias. To formalise the approach in a streamlined setting, we introduce a simple language to represent discontinuous functions piecewisely via if-statements, and we show how to systematically obtain a smoothed interpretation of such representations via sigmoid functions, which are parameterised by an *accuracy coefficient*.

Contributions. Our main contribution is the *provable* correctness of a novel variant of SGD, *Diagonalisation Stochastic Gradient Descent* (DSGD), to stationary points. The method takes gradient steps of



(a) Solid red: biased estimator $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\nabla_\theta f(\theta, z)\right]$, solid green: true gradient $\nabla_\theta\, \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[f(\theta, z)\right]$, black: gradient of smoothed objective (dotted: $\eta = 1$, dashed: $\eta = 1/3$) for Example 1.1.

(b) Sigmoid function σ_η (black dotted: $\eta = 1/3$, black dashed: $\eta = 1/15$) and the Heaviside step function (red, solid).

Figure 1

smoothed models whilst simultaneously enhancing the accuracy of the approximation in each iteration. Crucially, asymptotic correctness is not affected by the choice of (accuracy) hyperparameters.

We identify mild conditions on our language and the distribution for theoretical guarantees. In particular, for the smoothed problems we obtain unbiased gradient estimators, which converge uniformly to the true gradient as the accuracy is improved. Besides, as important ingredients for the correctness of DSGD we prove bounds on the variance, which solely depend on the syntactical structure of models.

Empirical studies show that DSGD performs comparably to the unbiased correction of the reparameterised gradient estimator by Lee et al. (2018). However our estimator is simpler, faster, and attains orders of magnitude reduction in work-normalised variance. Besides, DSGD exhibits more stable convergence than using an optimisation procedure for *fixed* accuracy coefficients (Khajwal et al., 2023), which is heavily affected by the choice of that accuracy coefficient.

Related Work. Lee et al. (2018) is the starting point for our work and a natural source for comparison. They correct the (biased) reparameterisation gradient estimator for non-differentiable models by additional non-trivial *boundary* terms. They present an efficient (but non-trivial) method for *affine* guards only. Besides, they are not concerned with the *convergence* of gradient-based optimisation procedures.

Maddison et al. (2017); Jang et al. (2017) study the reparameterisation gradient estimator and discontinuities arising from *discrete* random variables, and they propose a continuous relaxation. This can be viewed as a special case of our setting since discrete random variables can be encoded via continuous random variables and if-statements. In the context of discrete random variables, Tucker et al. (2017) combine the score estimator with control variates (a common variance reduction technique) based on such continuous approximations.

In practice, non-differentiable functions are often approximated smoothly. Some foundations of a similar smoothing approach are studied in (Zang, 1981, Thm. 3.1) in a non-stochastic setting.

Abstractly, our diagonalisation approach resembles graduated optimisation (Blake and Zisserman, 1987; Hazan et al., 2016): a “hard” problem is solved by “simpler” approximations in such a way that the quality of approximation improves over time. However, the goals and merits of the approaches are incomparable: graduated optimisation is concerned with overcoming non-convexity of the objective function to find global optima rather than stationary points, whereas our approach is motivated by overcoming the bias of the gradient estimator of the objective function.

Khajwal et al. (2023) study a (higher-order) probabilistic programming language and employ stochastic gradient descent on a fixed smooth approximation. Our DSGD algorithm advances their work in that it converges to stationary points of the *original* (unsmoothed) problem. Crucially, the accuracy coefficient does not need to be fixed in advance; rather it is progressively enhanced during the optimisation (which has important advantages such as higher robustness).

2 PROBLEM SETUP

We start by introducing a simple function calculus to represent (discontinuous) functions in a piecewise manner.

Let Op be a set of primitive functions/operations (which we restrict below) and $z_1, \ldots, z_n$ variables (for a fixed arity n). We define a class Expr of (syntactic) representations of piecewisely defined functions $\mathbb{R}^n \to \mathbb{R}$ inductively:

$$\text{Expr} \ni F ::= z_j \mid f(F, \ldots, F) \mid \mathbf{if}\ F < 0\ \mathbf{then}\ F\ \mathbf{else}\ F$$

where $f \in \text{Op}$. That is: expressions are nestings of the following ingredients: variables, function applications and if-conditionals. To enhance readability we use post- and infix notation for standard operations such as $+, \cdot, ^2$.

Example 2.1. Example 1.1 can be expressed as F_1:

$$\begin{aligned}
&F_1' \equiv \mathbf{if}\ z < 0\ \mathbf{then}\ 0\ \mathbf{else}\ 1\\
&F_1 \equiv -0.5 \cdot z^2 + \mathbf{if}\ z < 0\ \mathbf{then}\ 0\ \mathbf{else}\ 1\\
&F_2 \equiv \mathbf{if}\ (a \cdot (\mathbf{if}\ b \cdot z_1 + c < 0\ \mathbf{then}\ 0\ \mathbf{else}\ 1)\\
&\qquad\quad + d \cdot (\mathbf{if}\ e \cdot z_2 + f < 0\ \mathbf{then}\ 0\ \mathbf{else}\ 1) + g) < 0\\
&\qquad \mathbf{then}\ 0\ \mathbf{else}\ 1
\end{aligned}$$

F_2 illustrates that nested (if-)branching can occur not only in branches but also in the guard/condition. Such nestings arise in practice (see `xornet` in Section 6) and facilitates writing concise models.

$F \in \text{Expr}$ naturally defines (see Appendix A) a function $\mathbb{R}^n \to \mathbb{R}$, which we denote by $[\![F]\!]$. In particular, for f and F_1 defined in Examples 1.1 and 2.1, respectively, $[\![F_1]\!] = f$.

Expressivity. Our ultimate goal is to improve the inference engines of probabilistic programming languages such as Pyro (Bingham et al., 2019), which is built on Python and PyTorch and primarily uses variational inference. A crucial ingredient of such inference engines are low-variance, unbiased gradient estimators. We identify if-statements (which break continuity and differentiability) as the key challenge. We deliberately omit most features of mainstream languages because they are not relevant for the essence of the challenge and would only make the presentation a lot more complicated. In practice more language features will be desirable and it is worthwhile future work to extend the supported language.

Problem Statement. We are ready to formally state the problem we are solving in the present paper:

$$\text{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\mathbf{s} \sim \mathcal{D}}[[\![F]\!](\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))] \qquad (\mathbf{P})$$

where $F \in \text{Expr}$, $\mathcal{D}$ is a continuous probability distribution with support $\mathbb{S} \subseteq \mathbb{R}^n$, $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is the *parameter space* and each $\boldsymbol{\phi}_{\boldsymbol{\theta}} : \mathbb{S} \to \mathbb{S}$ is a diffeomorphism[a].

[a] i.e. a bijective differentiable function with differentiable inverse

Note that for $\mathcal{D}_{\boldsymbol{\theta}}(\mathbf{z}) := \mathcal{D}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})|$,

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!](\mathbf{z})] = \mathbb{E}_{\mathbf{s} \sim \mathcal{D}}[[\![F]\!](\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$$

Without further restrictions it is not *a priori* clear that the optimisation problem is well-defined due to a potential failure of integrability. Issues may be caused by both the distribution $\mathcal{D}$ and the expression F. E.g. the Cauchy distribution does not even have expectations, and despite $\mathcal{N}(\theta, 1)$—the normal distribution with mean θ and variance 1—being very well behaved, $\mathbb{E}_{z\sim\mathcal{N}(\theta,1)}[\exp(z^2)] = \infty$ regardless of θ.

Schwartz Functions. We slightly generalise the well-behaved class of Schwartz functions (see also for more details e.g. (Hörmander, 2015; Reed and Simon, 2003)) to accommodate probability density functions the support of which is a subset of $\mathbb{R}^n$:

A function $f : \mathbb{S} \to \mathbb{R}$, where $\mathbb{S} \subseteq \mathbb{R}^n$ is measurable and has measure-0 boundary, is a (generalised) *Schwartz function* if f is smooth in the interior of $\mathbb{S}$ and for all α and β (using standard multi-index notation for higher-order partial derivatives),

$$\sup_{\mathbf{x}\in\mathring{\mathbb{S}}} |\mathbf{x}^\beta \cdot \partial^\alpha f(\mathbf{x})| < \infty$$

Intuitively, a Schwartz function decreases rapidly.

Example 2.2. Distributions with pdfs which are also Schwartz functions include (for a fixed parameter) the (half) normal, exponential and logistic distributions.

Non-examples include the Cauchy distribution and the Gamma distributions. (The Gamma distribution cannot be reparameterised (Ruiz et al., 2016) and therefore it is only of marginal interest for our work regarding the reparameterisation gradient.)

The following pleasing properties of Schwartz functions (Hörmander, 2015; Reed and Simon, 2003) carry immediately over:

Lemma 2.3. *Let $f : \mathbb{S} \to \mathbb{R}$ be a Schwartz function.*

1. *All partial derivatives of f are Schwartz functions.*
2. *$f \in L^p(\mathbb{S})$; in particular f is integrable: $\int_{\mathbb{S}} |f(\mathbf{x})|\,\mathrm{d}\mathbf{x} < \infty$.*
3. *The product $(f \cdot p) : \mathbb{S} \to \mathbb{R}$ is also a Schwartz function if $p : \mathbb{R}^n \to \mathbb{R}$ is a polynomial.*

To mitigate the above (well-definedness) problem, we henceforth assume:

Assumption 2.4. 1. The density $\mathcal{D}$ is a (generalised) Schwartz function on its support $\mathbb{S} \subseteq \mathbb{R}^n$.
2. Op is the set of smooth functions all partial derivatives of which are bounded by polynomials.
3. $\boldsymbol{\phi}_{(-)}$ and its partial derivatives are bounded by polynomials and each $\boldsymbol{\phi}_{\boldsymbol{\theta}} : \mathbb{S} \to \mathbb{S}$ is a diffeomorphism.

This set-up covers in particular typical variational inference problems (see Section 6) with normal distributions because log-densities $f(x) \coloneqq \log(\mathcal{N}(x \mid \mu, \sigma^2))$ can be admitted as primitive operations.

Popular non-smooth functions such as ReLU, max or the absolute value function are piecewise smooth. Hence, they can be expressed in our language using if-statements and smooth primitives. For instance, wherever a user may wish to use $\mathrm{ReLU}(x)$, this can be replaced with the expression $\mathbf{if}\, x < 0\, \mathbf{then}\, 0\, \mathbf{else}\, x$.

Employing Lemma 2.3 we conclude that the objective function in (**P**) is well defined: for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!](\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))] < \infty$.

Whilst for this result it would have been sufficient to assume that just $\boldsymbol{\phi}_{(-)}$ and $f \in \mathrm{Op}$ (and not necessarily their derivatives) are polynomially bounded, this will become useful later (Section 5).

3 SMOOTHING

The bias of the reparametrisation gradient (cf. Example 1.1) is caused by discontinuities, which arise when interpreting if-statements in a standard way. Khajwal et al. (2023) instead avoid this problem by replacing the Heaviside step functions used in standard interpretations of if-statements with smooth approximations.

Formally, for $F \in \mathrm{Expr}$ and *accuracy coefficient* $\eta > 0$ we define the *η-smoothing* $[\![F]\!]_\eta : \mathbb{R}^n \to \mathbb{R}$:

$$\begin{aligned}
[\![z_j]\!]_\eta(\mathbf{z}) &\coloneqq z_j \\
[\![f(F_1,\dots,F_k)]\!]_\eta(\mathbf{z}) &\coloneqq f([\![F_1]\!]_\eta(\mathbf{z}),\dots,[\![F_k]\!]_\eta(\mathbf{z})) \\
[\![\mathbf{if}\, F < 0\, \mathbf{then}\, G\, \mathbf{else}\, H]\!]_\eta(\mathbf{z}) &\coloneqq \\
\sigma_\eta(-[\![F]\!]_\eta(\mathbf{z})) \cdot [\![G]\!]_\eta(\mathbf{z}) &+ \sigma_\eta([\![F]\!]_\eta(\mathbf{z})) \cdot [\![H]\!]_\eta(\mathbf{z})
\end{aligned}$$

where $\sigma_\eta(x) \coloneqq \sigma(\frac{x}{\eta}) = \frac{1}{1+\exp(-\frac{x}{\eta})}$ is the logistic sigmoid function (see Fig. 1b).

Note that the smoothing depends on the representation. In particular, $[\![F]\!] = [\![G]\!]$ does not necessarily imply $[\![F]\!]_\eta = [\![G]\!]_\eta$, e.g. $[\![\mathbf{if}\, z^2 < 0\, \mathbf{then}\, 0\, \mathbf{else}\, z]\!] = [\![z]\!]$ but $[\![\mathbf{if}\, z^2 < 0\, \mathbf{then}\, 0\, \mathbf{else}\, z]\!]_\eta \neq [\![z]\!]_\eta$.

Unbiasedness and SGD for Fixed Accuracy Coefficient. Each $[\![F]\!]_\eta$ is clearly differentiable. Therefore, the following is a consequence of a well-known result about exchanging differentiation and integration, which relies on the dominated convergence theorem (Khajwal et al., 2023; Klenke, 2014, Theorem 6.28):

Proposition 3.1 (Unbiasedness)**.** *For every $\eta > 0$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))] = \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[\nabla_{\boldsymbol{\theta}}\, [\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$$

Consequently, SGD can be employed on an η-smoothing for a fixed accuracy coefficient η.

Choice of Accuracy Coefficients. A natural question to ask is: *how do we choose an accuracy coefficient such that SGD solves the original, unsmoothed problem* (**P**) *"well"?* For our running example we can observe that this really matters: stationary points for low accuracy (i.e. high η) may not yield significantly better results than the biased (standard) reparameterisation gradient estimator (see Fig. 1a).

On the other hand, there is unfortunately no bound (as $\eta \searrow 0$) to the derivative of σ_η at 0 (see Fig. 1b). Therefore, the variance of the smoothed estimator also increases as the accuracy is enhanced.

In the following section we offer a principled solution to this problem with strong theoretical guarantees.

4 DIAGONALISATION SGD

We propose a novel variant of SGD in which we enhance the accuracy coefficient *during* optimisation (rather than *fixing* it in advance). For an expression $F \in \mathrm{Expr}$ and a sequence $(\eta_k)_{k\in\mathbb{N}}$ of accuracy coefficients we modify the standard SGD iteration to

$$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \gamma_k \nabla_{\boldsymbol{\theta}} [\![F]\!]_{\eta_k}(\boldsymbol{\theta}_k, \mathbf{s}_k) \qquad \mathbf{s}_k \sim \mathcal{D}$$

where γ_k is the step size. The qualifier "diagonal" highlights that, in contrast to standard SGD, we are not using the gradient of the same function $[\![F]\!]_\eta$ for each step but rather we are using the gradient of $[\![F]\!]_{\eta_k}$. Intuitively, this scheme facilitates getting close to the optimum whilst the variance is low (but the approximation may be coarse) and make small adjustments once the accuracy has been enhanced and approximation errors become visible.

Whilst the modification to the algorithm is moderate, we will be able to *provably guarantee* that asymptotically, the gradient of the original *unsmoothed* objective function vanishes.

To formalise the correctness result, we generalise the setting: suppose for each $k \in \mathbb{N}$, $f_k : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$ is differentiable. We define a *Diagonalisation Stochastic Gradient Descent* (DSGD) sequence:

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

Due to the aforementioned fact that also the variance increases as the accuracy is enhanced, the scheme of accuracy coefficients $(\eta_k)_{k\in\mathbb{N}}$ needs to be adjusted carefully to tame the growth of the variances V_k of the gradient of f_k, as stipulated by following equation:

$$\sum_{k\in\mathbb{N}} \gamma_k = \infty \qquad\qquad \sum_{k\in\mathbb{N}} \gamma_k^2 \cdot V_k < \infty \qquad (2)$$

In the regime $f_k = f$ and $V_k = V$ of standard SGD, this condition subsumes the classic condition by Robbins and Monro (1951), and $\gamma_k \in \Theta(1/k)$ is admissible.

The following exploits Taylor's theorem and can be obtained by modifying convergence proofs of standard SGD (see e.g. (Bertsekas and Tsitsiklis, 2000) or (Bertsekas, 2015, Chapter 2)):

Proposition 4.1 (Correctness)**.** *Suppose $(\gamma_k)_{k\in\mathbb{N}}$ and $(V_k)_{k\in\mathbb{N}}$ satisfy Eq. (2), $g_k(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}}[f_k(\boldsymbol{\theta}, \mathbf{s})]$ and $g(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}}[f(\boldsymbol{\theta}, \mathbf{s})]$ are well-defined and differentiable,*

(D1) $\nabla_{\boldsymbol{\theta}} g_k(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}}[\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{s})]$ for all $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
(D2) g is bounded, Lipschitz continuous and Lipschitz smooth[2] on $\boldsymbol{\Theta}$
(D3) $\mathbb{E}_{\mathbf{s}}[\|\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{s})\|^2] < V_k$ for all $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
(D4) ∇g_k converges uniformly to ∇g on $\boldsymbol{\Theta}$

Then almost surely[3] $\liminf_{i\to\infty} \|\nabla g(\boldsymbol{\theta}_i)\| = 0$ or $\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}$ for some $i \in \mathbb{N}$.

Having already discussed unbiasedness (D1), Proposition 3.1, we address the remaining premises in the next section to show that DSGD is correct for expressions.

5 ESTABLISHING PRE-CONDITIONS

Note that the pre-conditions of Proposition 4.1 may fail for non-compact $\boldsymbol{\Theta}$: e.g. the objective function $\theta \mapsto \mathbb{E}_{s\sim\mathcal{N}(0,1)}[s + \theta]$ is unbounded. Therefore, we assume the following henceforth:

Assumption 5.1. $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is compact.

For Lipschitz continuity it suffices to bound the partial derivatives of the objective function. Thus, we exchange differentiation and integration[4]:

$$\left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!](\mathbf{z})]\right| \le \int \left|\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{\partial}{\partial\theta_i}\mathcal{D}_{\boldsymbol{\theta}}(\mathbf{z})\right| \cdot |[\![F]\!](\mathbf{z})| \, \mathrm{d}\mathbf{z}$$

Extending the integrability result for Schwartz functions and polynomials (Lemma 2.3) in a non-trivial way (see Appendix D), we can demonstrate that the integral on the right side is finite. Our proof (cf. Appendix D.1) relies on the following:

Assumption 5.2. $\boldsymbol{\phi}_{(-)} : \boldsymbol{\Theta} \times \mathbb{S} \to \mathbb{S}$ satisfies

$$\inf_{(\boldsymbol{\theta},\mathbf{s})\in\boldsymbol{\Theta}\times\mathbb{S}} |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})| > 0$$

This requirement is a bit stronger than $|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})| > 0$ for each $(\boldsymbol{\theta}, \mathbf{s}) \in \boldsymbol{\Theta} \times \mathbb{S}$, which automatically holds if each $\boldsymbol{\phi}_{\boldsymbol{\theta}} : \mathbb{S} \to \mathbb{S}$ is a diffeomorphism.

Our prime examples, location-scale transformations, satisfiy this stronger property:

[2]Recall that a differentiable function $g : \boldsymbol{\Theta} \to \mathbb{R}$ is L-Lipschitz smooth if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\|\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}')\| \le L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

[3]w.r.t. the random choices $\mathbf{s}_1, \mathbf{s}_2, \ldots$ of DSGD

[4]This is valid because $\mathcal{D}_{\boldsymbol{\theta}}$ is differentiable and $[\![F]\!]$ is independent of θ_i.

Example 5.3. Suppose $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is compact, $f : \boldsymbol{\Theta} \to \mathbb{R}_{>0}$ is continuous and $g : \boldsymbol{\Theta} \to \mathbb{R}$. Then the location-scale transformation $\phi_{\boldsymbol{\theta}}(s) \coloneqq f(\boldsymbol{\theta}) \cdot s + g(\boldsymbol{\theta})$ satisfies Assumption 5.2 because $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta}) > 0$ ($\boldsymbol{\Theta}$ is compact and f is continuous).

Similarly, we can prove the other obligations of (D2).

5.1 Uniform Convergence of Gradients

Khajwal et al. (2023) show that under mild conditions, the smoothed objective function converges uniformly to the original, unsmoothed objective function. For (D4) we need to extend the result to gradients.

Recall that σ_η converges (pointwisely) to the Heaviside function on $\mathbb{R} \setminus \{0\}$ (cf. Fig. 1b). On the other hand, $\mathbb{E}_{z\sim\mathcal{N}(\theta,1)}[[\![\mathbf{if}\, 0 < 0\, \mathbf{then}\, 0\, \mathbf{else}\, 1]\!]_\eta(z)] = \frac{1}{2}$ converges *nowhere* to $\mathbb{E}_{z\sim\mathcal{N}(\theta,1)}[[\![\mathbf{if}\, 0 < 0\, \mathbf{then}\, 0\, \mathbf{else}\, 1]\!](z)] = 1$.

To rule out such contrived examples, we require that conditions in if-statements only use functions which are a.e. not 0. Formally, we define *safe* guards and expressions inductively by:

$$\mathrm{SGuard} \ni G ::= f(z_{i_1}, \ldots, z_{i_k}) \mid \mathbf{if}\, G < 0\, \mathbf{then}\, G\, \mathbf{else}\, G$$
$$\mathrm{SExpr} \ni S ::= z_j \mid f(S, \ldots, S) \mid \mathbf{if}\, G < 0\, \mathbf{then}\, S\, \mathbf{else}\, S$$

where in the first rule we assume that $f \neq 0$ a.e.[5] and the i_j are pairwise distinct.

Note that SGuard $\subset$ SExpr $\subset$ Expr and for $G \in$ SGuard, $[\![G]\!] \neq 0$ a.e. As a consequence, by structural induction, we can show that for $S \in$ SExpr, $[\![S]\!]_\eta$ converges to $[\![S]\!]$ almost everywhere.

Exploiting a.e. convergence we conclude not only the uniform convergence of the smoothed objective function but also their *gradients*:

Proposition 5.4 (Uniform Convergence)**.** *If* $F \in$ SExpr *then*

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!]_\eta(\mathbf{z})] \xrightarrow{\text{unif.}} \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!](\mathbf{z})]$$
$$\nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!]_\eta(\mathbf{z})] \xrightarrow{\text{unif.}} \nabla_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!](\mathbf{z})]$$

as $\eta \searrow 0$ *for* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

5.2 Bounding the Variance

Next, we analyse the variance for (D3). Recall that nesting if-statements in guards (e.g. F_2 in Example 2.1) results in nestings of σ_η in the smoothed interpretation, which in view of the chain rule may cause high variance as $\sigma'_\eta(0) = \frac{1}{4\eta}$. Therefore, to give good bounds, we classify expressions by their maximal nesting depth ℓ of the conditions of if-statements. Formally, we define Expr_ℓ inductively

1. $z_j \in \mathrm{Expr}_0$
2. If $f : \mathbb{R}^k \to \mathbb{R} \in \mathrm{Op}$ and $F_1, \ldots, F_k \in \mathrm{Expr}_\ell$ then $f(F_1, \ldots, F_k) \in \mathrm{Expr}_\ell$
3. If $F \in \mathrm{Expr}_\ell$ and $G, H \in \mathrm{Expr}_{\ell+1}$ then $(\mathbf{if}\, F < 0\, \mathbf{then}\, G\, \mathbf{else}\, H) \in \mathrm{Expr}_{\ell+1}$
4. If $F \in \mathrm{Expr}_\ell$ then $F \in \mathrm{Expr}_{\ell+1}$

Note that $\mathrm{Expr} = \bigcup_{\ell\in\mathbb{N}} \mathrm{Expr}_\ell$. For the expressions in Example 2.1, $F_1, F'_1 \in \mathrm{Expr}_1$ and $F_2 \in \mathrm{Expr}_2$.

Now, exploiting the chain rule and the fact that $|\sigma'_\eta| \leq \eta^{-1}$, it is relatively straightforward to show inductively that for $F \in \mathrm{Expr}_\ell$ there exists $L > 0$ such that $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[\|\nabla_{\boldsymbol{\theta}}[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))\|^2] \leq L \cdot \eta^{-2\ell}$ for all $\eta > 0$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. However, we can give a sharper bound, which will allows us to enhance the accuracy more rapidly (in view of Eq. (2)):

Proposition 5.5. *If* $F \in \mathrm{Expr}_\ell$ *then there exists* $L > 0$ *such that for all* $\eta > 0$ *and* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[\|\nabla_{\boldsymbol{\theta}}[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))\|^2] \leq L \cdot \eta^{-\ell}$.

To get an intuition of the proof presented in Appendix D.2, we consider our running Example 2.1, $F'_1 \equiv (\mathbf{if}\, z < 0\, \mathbf{then}\, 0\, \mathbf{else}\, 1) \in \mathrm{Expr}_1$ and $\phi_\theta(s) = s + \theta$:

$$\begin{aligned}
&\mathbb{E}_{s\sim\mathcal{N}}\left[\left|\frac{\partial([\![F'_1]\!]_\eta \circ \phi_{(-)})}{\partial\theta}(\theta, s)\right|^2\right] = \mathbb{E}\left[\left(\sigma'_\eta(\phi_\theta(s))\right)^2\right] \\
&\leq \eta^{-1} \cdot \int \mathcal{N}(s) \cdot \frac{\partial(\sigma_\eta \circ \phi_{(-)})}{\partial s}(\theta, s)\,\mathrm{d}s \\
&= \eta^{-1} \cdot \left([\mathcal{N}(s) \cdot \sigma_\eta(\phi_\theta(s))]_{-\infty}^{\infty} - \int \mathcal{N}'(s) \cdot \sigma_\eta(\phi_\theta(s))\,\mathrm{d}s\right) \\
&\leq \eta^{-1} \cdot \int |\underbrace{\mathcal{N}'(s)}_{\text{Schwartz}}|\,\mathrm{d}s \leq \eta^{-1} \cdot L
\end{aligned}$$

where we used integration by parts in the third step.

5.3 Concluding Correctness

Having bounded the variance, we present a scheme of accuracy coefficients compatible with the scheme of step sizes $\gamma_k = 1/k$, which is the classic choice for SGD. Note that for any $\epsilon > 0$, $\sum_{k\in\mathbb{N}} \frac{1}{k^2} \cdot (k^{\frac{1}{\ell}-\epsilon})^\ell < \infty$. Therefore, by Proposition 5.5 we can choose accuracy coefficients $\eta_k \in \Theta(k^{-\frac{1}{\ell}+\epsilon})$. Finally, with Propositions 4.1 and 5.4 we conclude the correctness of DSGD for smoothings:

Theorem 5.6 (Correctness of DSGD)**.** *Let* $F \in \mathrm{Expr}_\ell \cap \mathrm{SExpr}$ *and* $\epsilon > 0$. *Then DSGD is correct for* $[\![F]\!]_{\eta_k}$, $\gamma_k \in \Theta(1/k)$ *and* $\eta_k \in \Theta(k^{-\frac{1}{\ell}+\epsilon})$*:*

[5] By Mityagin (2015) this can be guaranteed if f is analytic, not constantly 0. (Recall that a function $f : R^m \to R^n$ is analytic if it is infinitely differentiable and its multivariate Taylor expansion at every point $x_0 \in \mathbb{R}^m$ converges pointwise to f in a neighbourhood of x_0.)

almost surely $\liminf_{i\to\infty} \|\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!](\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]\| = 0$ *or* $\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}$ *for some* $i \in \mathbb{N}$.

For instance for $F \in \mathrm{Expr}_1$ we can choose $\eta_k \in \Theta(1/\sqrt{k})$. Crucially, for the choice of accuracy coefficients only the syntactic structure (i.e. nesting depth) of terms is essential. In particular, there is no need to calculate bounds on the Hessian, the constant in Proposition 5.5, etc.

5.3.1 Discussion

Accuracy coefficient schedule for other learning rate schemes. The requirement on the step sizes and accuracy coefficients stipulated by Eq. (2) can be relaxed to

$$(\sum_{k\in\mathbb{N}} \gamma_k)^{-1} \cdot \sum_{k\in\mathbb{N}} \gamma_k^2 \cdot V_k = 0 \tag{3}$$

which is useful for deriving other admissible accuracy coefficient schemes. For instance, step size $\gamma_k = 1/\sqrt{k}$ violates Eq. (2). However, for terms with nesting depth $\ell = 1$, the accuracy coefficients $\eta_k = k^{-0.5+\epsilon}$ (where $\epsilon > 0$ is arbitrary) satisfy this relaxed requirement, and we obtain the same correctness guarantees as Theorem 5.4.

Choice of ϵ. Whilst asymptotically *any* choice of $\epsilon > 0$ enjoys the theoretical guarantees of Theorem 5.6, the *speed* of convergence of the gradient norm of the unsmoothed objective is governed by two summands: (1) the speed of convergence of the gradient of the smoothed objective function to the gradient of the unsmoothed objective, and (2) the accumulated contribution of the variances (more formally: the speed of convergence of the finite sums of the relaxation Eq. (3) above of Eq. (2) to 0 and thus the magnitude of $\sum_{i\in\mathbb{N}} 1/k^2 \cdot (k^{1/\ell-\epsilon})^{\ell}$).

Whilst for the former, small $\epsilon > 0$ is beneficial (enhancing the accuracy rapidly), for the latter large ϵ is beneficial. Consequently, for best performance in practice a trade-off between the two effects needs to be made, and we found the choices reported below to work well.

Average Variance of Run By Proposition 5.5 the average variance of a finite DSGD run with length N is bounded by (using Hölder's inequality)

$$\frac{1}{N} \cdot \sum_{k=1}^{N} \eta_k^{-\ell} \leq \eta_{\frac{N+1}{2}}^{-\ell}$$

Consequently, the average variance of a DSGD run is lower than for standard SGD with a fixed accuracy coefficient $\eta < \eta_{\frac{N+1}{2}}$, the accuracy coefficient of DSGD after half the iterations.

6 EMPIRICAL EVALUATION

We evaluate our DSGD procedure against SGD with the following gradient estimators: the biased reparameterisation estimator (Reparam), the unbiased correction thereof (LYY18, (Lee et al., 2018)), the smoothed reparametrisation estimator of Khajwal et al. (2023) for a *fixed* accuracy coefficient (Fixed), and the unbiased (Score) estimator.

Models. We include the models from Lee et al. (2018); Khajwal et al. (2023) and add a `random-walk` model. We summarise some details (the additional models are covered in Appendix E):

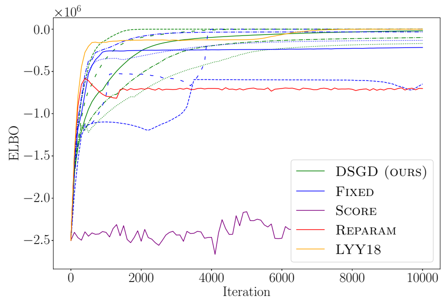
`temperature` Soudjani et al. (2017) model a controller keeping the temperature of a room within set bounds. The discontinuity arises from the discrete state of the controller, being either on or off. The model has a 41-dimensional latent variable and 80 if-statements.

`random-walk` models a random walk (similar to (Mak et al., 2021)) of bounded length. The goal is to infer the starting position based on the distance walked. The walk stops as soon as the destination is reached. This is checked using if-statements and causes discontinuities. In each step a normal-distributed step is sampled and its absolute value is added to the distance walked so far, which accounts for more non-differentiabilities. Overall, the model has a 16-dimensional latent variable and 31 if-statements.

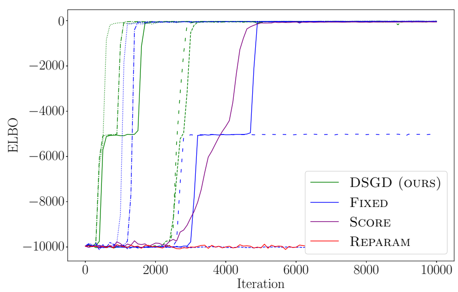
`xornet` is a multi-layer neural network trained to compute the XOR function with all activation functions being the Heaviside step function. The model has a 25-dimensional latent space (for all the weights and biases) and 28 if-statements. The Lyy18 estimator is *not* applicable to this model since the branch conditions are not all affine in the latent space.

F_2 from Example 2.1 can be viewed as a (stark) simplification of `xornet`. It is also the only model in which the guards of if-statements contain variables which in turn depend on branching. As such, `xornet` corresponds to a term in Expr_3, whereas all other models correspond to Expr_1.

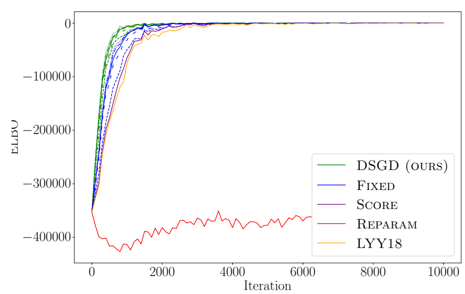
Experimental Set-Up. The (Python) implementation is based on (Khajwal et al., 2023; Lee et al., 2018). We employ the `jax` library to provide automatic differentiation which is used to implement each of the above estimators for an arbitrary (probabilistic) program. The smoothed interpretation can be obtained automatically by (recursively) replacing conditionals **if** $E_1 < 0$ **then** E_2 **else** E_3 with $\sigma_\eta(-E_1)\cdot E_2 + \sigma_\eta(E_1)\cdot E_3$ in a preprocessing step. (We avoid a potential blowup by using an auxiliary variable for E_1.)



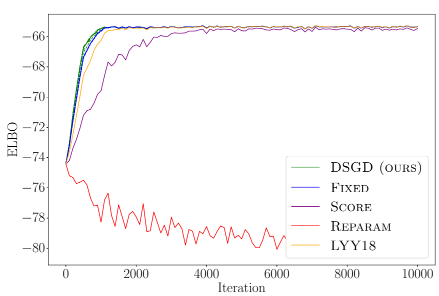
(a) `temperature`



(b) `xornet`



(c) `random-walk`



(d) `cheating`

Figure 2: ELBO trajectories for each model. A single colour is used for each estimator and the choice of $\eta = \eta_{4000} = 0.06, 0.1, 0.14, 0.18, 0.22$ (which determines η_0) is represented by dashed, loosely dashed, solid, dash-dotted, dotted lines, respectively.

In view of Theorem 5.6, for DSGD we choose the accuracy coefficient schemes $\eta_k := \eta_0/\sqrt{k}$ for $\eta_0 > 0$; due to the nesting of guards we use $\eta_k := \eta_0 \cdot k^{-0.2}$ for `xornet`. We compare (using the same line style) DSGD for different choices[6] of η_0 to Fixed using the fixed accuracy coefficient corresponding to η_{4000}.

To enable a fair comparison to (Khajwal et al., 2023; Lee et al., 2018), we follow their set-up and use the state-of-the-art stochastic optimiser `Adam`[7] with a step size of `0.001`, except for `xornet` for which we use `0.01`, for 10,000 iterations. For each iteration, we use 16 Monte Carlo samples from the chosen estimator to compute the gradient. As in (Lee et al., 2018), the LYY18 estimator does not compute the boundary surface term exactly, but estimates it using a single subsample.

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 are the product of the computational cost and the variance[8].

Analysis of Results. The ELBO trajectories as well as the data for computational cost and variance are presented Fig. 2 and Table 2 (additional models are covered in Appendix E). Besides, Table 1 lists the mean and standard deviation of the final ELBO across different seeds for the random number generator. Empirically, the bias of Reparam becomes evident and Score exhibits very high variance, resulting in slow convergence or even inferior results (Fig. 2a).

Whenever the LYY18 estimator is applicable[9], the trajectories for DSGD perform comparably, however DSGD attains orders of magnitude reduction in work-normalised variance (4 to 20,000 x).

Compared to Fixed, we observe that DSGD is more robust[10] to the choice of (initial) accuracy coefficients (especially for `temperature` and `xornet`). Besides,

[6]The choice of benchmarked hyperparameters for Fig. 2 expands the range by Khajwal et al. (2023), who use 0.1, 0.15, 0.2, and we uniformly split the range in steps of 0.04.

[7]together with the respective gradient estimators, e.g. $\nabla[\![F]\!]_{\eta_k}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))$ in step k for DSGD

[8]This is a more suitable measure than "raw" variances since the latter can be improved routinely at the expense of computational efficiency by taking more samples.

[9]For `xornet`, LYY18 is *not* applicable as there are non-affine conditions in if-statements.

[10]On a *finite* run our *asymptotic* convergence result Theorem 5.6 cannot completely eliminate the dependence on this choice.

Table 1: Mean of the final ELBO (the higher the better) for different random seeds and indicating error bars (the ± is one standard deviation).

(a) `temperature`

η/η_{4000}	DSGD (ours)	Fixed	Score	Reparam	LYY18
0.06	-76 ± 1	-624,250 ± 44,121			
0.1	-84 ± 2	-425 ± 9	-2,611,479 ± 255,193	-706,729 ± 4,697	-17,502 ± 52,044
0.14	-15,476 ± 4,641	-121,932 ± 85,460			

(b) `xornet`

η/η_{4000}	DSGD (ours)	Fixed	Score	Reparam
0.06	-3,530 ± 3,889	-5,522 ± 4,136		
0.1	-33 ± 7	-2,029 ± 3,305	-553 ± 1,507	-9,984 ± 38
0.14	-27 ± 4	-2,028 ± 3,986		

Table 2: 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($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	1.71	4.91e-11	2.54e-10
Fixed	1.71	2.84e-10	2.24e-09
Reparam	1.26	1.47e-08	1.94e-08
LYY18	9.61	1.05e-06	4.04e-05

(b) `xornet`

Estimator	Cost	Avg($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	1.74	6.21e-03	3.66e-02
Fixed	1.87	1.21e-02	5.43e-02
Reparam	0.388	8.34e-09	2.62e-09

(c) `random-walk`

Estimator	Cost	Avg($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	4.70	1.71e-01	2.61e-01
Fixed	4.70	9.50e-01	1.49
Reparam	2.17	8.63e-10	7.01e-10
LYY18	4.81	7.92	1.26e+01

(d) `cheating`

Estimator	Cost	Avg($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	1.52	2.31e-03	3.51e-03
Fixed	1.52	2.84e-03	4.64e-03
Reparam	9.36e-01	4.14e-19	1.16e-18
LYY18	2.59	4.27e-02	1.09e-01

there is a moderate improvement of variance (Table 2).

7 CONCLUDING REMARKS

We have proposed a variant of SGD, *Diagonalisation Stochastic Gradient Descent*, and shown *provable* correctness. Our approach is based on a smoothed interpretation of (possibly) discontinuous programs, which also yields unbiased gradient estimators. Crucially, asymptotically a stationary point of the original, *unsmoothed* problem is attained and a hyperparameter (accuracy of approximations) is tuned automatically. The correctness hinges on a careful analysis of the variance and a compatible scheme governing the accuracies. Notably, this purely depends on the (syntactic) structure of the program.

Our experimental evaluation demonstrates important advantages over the state of the art: significantly lower variance (score estimator), unbiasedness (reparametrisation estimator), simplicity, wider applicability and lower variance (unbiased correction thereof), as well as stability over the choice of (initial) accuracy coefficients (fixed smoothing).

Limitations and Future Directions. Our analysis is asymptotic and focuses on stationary points, which leaves room for future research (convergence rates, avoidance of saddle points etc.).

Furthermore, we plan to explore methods *adaptively* tuning the accuracy coefficient rather than *a priori* fixing a scheme. Whilst the present work was primarily concerned with theoretical guarantees, we anticipate adaptive methods to outperform fixed schemes in practice (but likely without theoretical guarantees).

Acknowledgements

This research is supported by the National Research Foundation, Singapore, under its RSS Scheme (NRF-RSS2022-009).

References

Barthe, G., Katoen, J.-P., and Silva, A., editors (2020). *Foundations of Probabilistic Programming*. Cambridge University Press.

Bertsekas, D. (2015). *Convex optimization algorithms*. Athena Scientific.

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

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6.

Blake, A. and Zisserman, A. (1987). *Visual Reconstruction*. MIT Press.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Botev, Z. and Ridder, A. (2017). Variance Reduction. In *Wiley StatsRef: Statistics Reference Online*, pages 1–6.

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

Fu, M. C. (2006). Chapter 19 gradient estimation. In Henderson, S. G. and Nelson, B. L., editors, *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, pages 575–616. North-Holland.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. (2016). On graduated optimization for stochastic non-convex problems. In Balcan, M. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1833–1841. JMLR.org.

Hörmander, L. (2015). *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Classics in Mathematics. Springer Berlin Heidelberg.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Khajwal, B., Ong, C. L., and Wagner, D. (2023). Fast and correct gradient-based optimisation for probabilistic programming via smoothing. In Wies, T., editor, *Programming Languages and Systems - 32nd European Symposium on Programming, ESOP 2023, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2023, Paris, France, April 22-27, 2023, Proceedings*, volume 13990 of *Lecture Notes in Computer Science*, pages 479–506. Springer.

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

Lee, W., Yu, H., and Yang, H. (2018). 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.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). 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*.

Mak, C., Ong, C. L., Paquet, H., and Wagner, D. (2021). Densities of almost surely terminating probabilistic programs are differentiable almost everywhere. In Yoshida, N., 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.

Minh, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. 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 1791–1799. JMLR.org.

Mityagin, B. (2015). The zero set of a real analytic function.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21:132:1–132:62.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). 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.

Reed, M. and Simon, B. (2003). *Methods of Modern Mathematical Physics: Functional analysis. I.* World Published Corporation.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). 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.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruiz, F. J. R., Titsias, M. K., and Blei, D. M. (2016). The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 460–468.

Schulman, J., Heess, N., Weber, T., and Abbeel, P. (2015). Gradient estimation using stochastic computation graphs. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3528–3536.

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

Soudjani, S. E. Z., Majumdar, R., and Nagapetyan, T. (2017). Multilevel monte carlo method for statistical model checking of hybrid systems. In Bertrand, N. and Bortolussi, L., 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.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tucker, G., Mnih, A., Maddison, C. J., Lawson, D., and Sohl-Dickstein, J. (2017). REBAR: low-variance, unbiased gradient estimates for discrete latent variable models. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2627–2636.

van de Meent, J., Paige, B., Yang, H., and Wood, F. (2018). An introduction to probabilistic programming. *CoRR*, abs/1809.10756.

Wingate, D. and Weber, T. (2013). Automated variational inference in probabilistic programming. *CoRR*, abs/1301.1299.

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. (2019). Variance reduction properties of the reparameterization trick. In Chaudhuri, K. and Sugiyama, M., editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 2711–2720. PMLR.

Zang, I. (1981). Discontinuous optimization by smoothing. *Mathematics of Operations Research*, 6(1):140–152.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):2008–2026.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   We present the setting in Section 2 and state Assumptions 2.4 and 5.1. The DSGD algorithm is presented in Section 4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   We prove an (asymptotic) convergence result Proposition 4.1 and Theorem 5.6.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]
   The Python code is at `https://github.com/domwagner/DSGD.git`*. The dependencies are* `jax`, `jaxlib`, `numpy`, `ipykernel`, `matplotlib`.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   See Assumptions 2.4 and 5.1.
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   Proofs are presented in Appendices B to D.
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

We discuss Assumptions 2.4 and 5.2 in Sections 2 and 5.1, respectively.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   The code is available at `https://github.com/domwagner/DSGD.git`.
   The experiements can be viewed and run in the jupyter notebook `experiments.ipynb` *by running:*
   `jupyter notebook experiments.ipynb`.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   We present the details Section 6 and Fig. 2 and closely follow the setup of Lee et al. (2018); Khajwal et al. (2023). In particular, the choice of the benchmarked hyperparameters $\eta = \eta_{4000} = 0.06, 0.1, 0.14, 0.18, 0.22$ *for Fig. 2 expands the range by Khajwal et al. (2023), who use* $0.1, 0.15, 0.2$*, and we uniformly split the range in steps of* 0.04.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   We describe our approach to compute the work-normalised variance and the computational cost in Section 6.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]
   We run our experiments on a MacBook Air (13-inch, 2017) with Intel HD Graphics 6000 1536 MB.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A Supplementary Materials for Section 2

$F \in \text{Expr}$ naturally defines a function $\mathbb{R}^n \to \mathbb{R}$, which we denote by $[\![F]\!]$:

$$\begin{aligned}[\![z_j]\!](\mathbf{z}) &\coloneqq z_j \\ [\![f(F_1,\dots,F_k)]\!](\mathbf{z}) &\coloneqq f([\![F_1]\!](\mathbf{z}),\dots,[\![F_k]\!](\mathbf{z})) \\ [\![\mathbf{if}\, F<0\,\mathbf{then}\, G\,\mathbf{else}\, H]\!](\mathbf{z}) &\coloneqq \begin{cases}[\![G]\!](\mathbf{z}) & \text{if } [\![F]\!](\mathbf{z})<0 \\ [\![H]\!](\mathbf{z}) & \text{otherwise}\end{cases}\end{aligned}$$

Table 3: Schwartz Distributions and their PDFs

distribution	pdf	support	parameters
normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{R}$	$\mu\in\mathbb{R},\sigma\in\mathbb{R}_{>0}$
half normal	$\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp\left(\frac{-x^2}{2\sigma^2}\right)$	$\mathbb{R}_{\geq 0}$	$\sigma\in\mathbb{R}_{>0}$
exponential	$\lambda\cdot\exp(-\lambda x)$	$\mathbb{R}_{\geq 0}$	$\lambda\in\mathbb{R}_{>0}$
logistic	$\frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}$	$\mathbb{R}$	$\mu\in\mathbb{R},s\in\mathbb{R}_{>0}$

B Supplementary Materials for Section 3

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.1. *Let $\Theta\subseteq\mathbb{R}$ be open and $\mathbb{S}\subseteq\mathbb{R}^n$ be measurable. If $f:\Theta\times\mathbb{S}\to\mathbb{R}$ satisfies*

1. *for each $\theta\in\Theta$, $\mathbf{s}\mapsto f(\theta,\mathbf{s})$ is integrable*
2. *f is differentiable*
3. *there exists an integrable $g:\mathbb{S}\to\mathbb{R}$ satisfying $|\frac{\partial f}{\partial\theta}(\theta,\mathbf{s})|\leq g(\mathbf{s})$ for all $(\theta,\mathbf{s})\in\Theta\times\mathbb{S}$.*

then for all $\theta\in\Theta$, $\frac{\partial}{\partial\theta}\int_{\mathbb{S}} f(\theta,\mathbf{s})\,\mathrm{d}\mathbf{s}=\int_{\mathbb{S}}\frac{\partial f}{\partial\theta}(\theta,\mathbf{s})\,\mathrm{d}\mathbf{s}$.

Note that the second premise fails for the function f in Example 1.1: for *all* $s\in\mathbb{R}$, $\frac{\partial f}{\partial\theta}(-s,s)$ does not exist.

Proposition 3.1 (Unbiasedness)**.** *For every $\eta>0$ and $\boldsymbol{\theta}\in\boldsymbol{\Theta}$,*

$$\nabla_{\boldsymbol{\theta}}\,\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]=\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[\nabla_{\boldsymbol{\theta}}\,[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$$

Proof. Let $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. We apply Lemma B.1 to some ball $\boldsymbol{\Theta}'\subseteq\boldsymbol{\Theta}$ around $\boldsymbol{\theta}$ and $(\mathbf{s},\boldsymbol{\theta})\mapsto\mathcal{D}(\mathbf{s})\cdot[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))$. We have already seen the well-definedness (first premise) and the second is obvious (since $[\![F]\!]_\eta$ is a smoothing). For the third premise, we observe that for each $\eta>0$, $\frac{\partial([\![F]\!]_\eta\circ\boldsymbol{\phi}_{(-)})}{\partial\theta_i}$ is bounded by a polynomial (using Assumption 2.4). Therefore, by Lemma B.2 (below) there exists a polynomial $p:\mathbb{R}^n\to\mathbb{R}$ satisfying $\frac{\partial([\![F]\!]_\eta\circ\boldsymbol{\phi}_{(-)})}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{s})\leq|p(\mathbf{s})|$ for all $(\boldsymbol{\theta},\mathbf{s})\in\boldsymbol{\Theta}'\times\mathbb{R}^n$ and integrability of $\mathcal{D}(\mathbf{s})\cdot p(\mathbf{s})$ follows with Lemma 2.3. $\square$

Lemma B.2. *1. If $p:\mathbb{R}^n\to\mathbb{R}$ is a polynomial then there exists a polynomial q such that for all $|x_1|\leq|x_1'|,\dots,|x_n|\leq|x_n'|$, $|p(\mathbf{x})|\leq q(\mathbf{x}')$.*

2. If $f:\boldsymbol{\Theta}\times\mathbb{R}^n\to\mathbb{R}$ is bounded by a polynomial, where $\boldsymbol{\Theta}\subseteq\mathbb{R}^m$ is compact then there exists a polynomial $p:\mathbb{R}^n\to\mathbb{R}$ satisfying $|f(\boldsymbol{\theta},\mathbf{x})|\leq p(\mathbf{x})$ for all $(\boldsymbol{\theta},\mathbf{x})\in\boldsymbol{\Theta}\times\mathbb{R}^n$.

For example for $p(x_1,x_2)=x_1^2\cdot x_2-x_2$ the polynomial $x_1^2\cdot(x_2^2+1)+(x_2^2+1)$ satisfies this property. (The following proof yields $((x_1^2+1)^2+1)\cdot(x_2^2+1)+(x_2^2+1)$.) Besides, $p(\theta,x_1,x_2)=\theta\cdot x_1\cdot x_2-x_2$ is uniformly bounded by $2\cdot(x_1^2+1)\cdot(x_2^2+1)+(x_2^2+1)$ on $\Theta=(-2,1)$.

Proof. 1. If $p(\mathbf{x})=x_i$ then we can choose $q(\mathbf{x})\coloneqq(x_i^2+1)$ because for $|x_i|\leq|x_i'|$, $|x_i|\leq|x_i'|<(x_i')^2+1$.

If $p(\mathbf{x}) = c$ for $c \in \mathbb{R}$ then we can choose $q(\mathbf{x}) := |c|$.
Finally, suppose that p_1, p_2, q_1, q_2 are polynomials such that for all $|\mathbf{x}| \leq |\mathbf{x}'|$, $|p_1(\mathbf{x})| \leq q_1(\mathbf{x}')$ and $|p_2(\mathbf{x})| \leq q_2(\mathbf{x}')$. Then for all $|\mathbf{x}| \leq |\mathbf{x}'|$,

$$|p_1(\mathbf{x}) + p_2(\mathbf{x})| \leq |p_1(\mathbf{x})| + |p_2(\mathbf{x})| \leq q_1(\mathbf{x}') + q_2(\mathbf{x}')$$
$$|p_1(\mathbf{x}) \cdot p_2(\mathbf{x})| \leq |p_1(\mathbf{x})| \cdot |p_2(\mathbf{x})| \leq q_1(\mathbf{x}') \cdot q_2(\mathbf{x}')$$

2. If $f : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$ is bounded by a polynomial then by the first part there exists $q : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$ such that for all $(\boldsymbol{\theta}, \mathbf{z}), (\boldsymbol{\theta}', \mathbf{z}) \in \boldsymbol{\Theta} \times \mathbb{R}^n$ with $|\boldsymbol{\theta}| \leq |\boldsymbol{\theta}'|$, $|f(\boldsymbol{\theta}, \mathbf{z})| \leq q(\boldsymbol{\theta}, \mathbf{z}')$. Let $\theta_i^* := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \theta_i < \infty$ ($\boldsymbol{\Theta}$ is bounded) and $p(\mathbf{z}) := q(\mathbf{z}, \boldsymbol{\theta}^*)$. Finally, it suffices to note that for every $(\boldsymbol{\theta}, \mathbf{z}) \in \boldsymbol{\Theta} \times \mathbb{R}^n$, $|f(\boldsymbol{\theta}, \mathbf{z})| \leq q(\boldsymbol{\theta}^*, \mathbf{z}) = p(\mathbf{z})$. $\square$

C Supplementary Materials for Section 4

Proposition 4.1 (Correctness)**.** *Suppose $(\gamma_k)_{k\in\mathbb{N}}$ and $(V_k)_{k\in\mathbb{N}}$ satisfy Eq. (2), $g_k(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}}[f_k(\boldsymbol{\theta}, \mathbf{s})]$ and $g(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}}[f(\boldsymbol{\theta}, \mathbf{s})]$ are well-defined and differentiable,*

(D1) $\nabla_{\boldsymbol{\theta}} g_k(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}}[\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{s})]$ for all $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
(D2) g is bounded, Lipschitz continuous and Lipschitz smooth[11] on $\boldsymbol{\Theta}$
(D3) $\mathbb{E}_{\mathbf{s}}[\|\nabla_{\boldsymbol{\theta}} f_k(\boldsymbol{\theta}, \mathbf{s})\|^2] < V_k$ for all $k \in \mathbb{N}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
(D4) ∇g_k converges uniformly to ∇g on $\boldsymbol{\Theta}$

Then almost surely[12] $\liminf_{i\to\infty} \|\nabla g(\boldsymbol{\theta}_i)\| = 0$ or $\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}$ for some $i \in \mathbb{N}$.

Proof. Let L be the constant for Lipschitz smoothness in (D2). By Taylor's theorem, $\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1} \in \boldsymbol{\Theta}$ and convexity of $\boldsymbol{\Theta}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{s}_k\sim\mathcal{D}}[g(\boldsymbol{\theta}_{k+1})] &\leq \mathbb{E}_{\mathbf{s}_k\sim\mathcal{D}}\left[g(\boldsymbol{\theta}_k) - \gamma_k \cdot \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_k), \nabla_{\boldsymbol{\theta}}\, f_k(\boldsymbol{\theta}_k, \mathbf{s}_k)\rangle + \frac{\gamma_k^2}{2} \cdot L \cdot \|\nabla_{\boldsymbol{\theta}}\, f_k(\boldsymbol{\theta}_k, \mathbf{s}_k)\|^2\right] \\ &\leq g(\boldsymbol{\theta}_k) - \gamma_k \cdot \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_k), \nabla_{\boldsymbol{\theta}}\, g_k(\boldsymbol{\theta}_k)\rangle + \frac{\gamma_k^2}{2} \cdot L \cdot V_k\end{aligned}$$

using (D1) and (D3) in the second step. Hence,

$$\gamma_k \cdot \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_k), \nabla_{\boldsymbol{\theta}}\, g_k(\boldsymbol{\theta}_k)\rangle \leq g(\boldsymbol{\theta}_k) - \mathbb{E}_{\mathbf{s}_k\sim\mathcal{D}}[g(\boldsymbol{\theta}_{k+1})] + \frac{\gamma_k^2}{2} \cdot L \cdot V_k$$

and thus,

$$\sum_{i=0}^{k-1} \gamma_i \cdot \mathbb{E}_{\mathbf{s}_0,\dots,\mathbf{s}_{k-1}\sim\mathcal{D}}[\langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle] \leq g(\boldsymbol{\theta}_0) - \mathbb{E}[g(\boldsymbol{\theta}_k)] + L \cdot \sum_{i=0}^{k-1} \frac{\gamma_i^2}{2} \cdot V_i$$

By boundedness (D2), $g(\boldsymbol{\theta}_0) - \mathbb{E}[g(\boldsymbol{\theta}_k)] \leq c < \infty$ (for some $c > 0$) and therefore due to Eq. (2) it follows:

$$\inf_{i\in\mathbb{N}} \mathbb{E}[\langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle] \leq \left(\sum_{i\in\mathbb{N}} \gamma_i\right)^{-1} \cdot \left(c + L \cdot \sum_{i\in\mathbb{N}} \frac{\gamma_i^2}{2} \cdot V_i\right) = 0 \tag{4}$$

By the same reasoning for all $i_0 \in \mathbb{N}$, $\inf_{i\geq i_0} \mathbb{E}[\langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle] = 0$ and therefore also $\liminf_{i\in\mathbb{N}} \mathbb{E}[\langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle] = 0$.

Next, observe that

$$\begin{aligned}\|g(\boldsymbol{\theta}_i)\|^2 &= \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle + \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle \\ &\leq \langle \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i), \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\rangle + \|\nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i)\| \cdot \|\nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\|\end{aligned}$$

and the right summand uniformly converges to 0 because of Lipschitz continuity (D2) and the fact that $\|\nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\theta}}\, g_i(\boldsymbol{\theta}_i)\|$ does so (D4). Consequently,

$$\liminf_{i\in\mathbb{N}} \mathbb{E}\left[\|g(\boldsymbol{\theta}_i)\|^2\right] = 0$$

Finally, this can only possibly hold if almost surely $\liminf_{i\in\mathbb{N}} \|g(\boldsymbol{\theta}_i)\|^2 = 0$. $\square$

[11] Recall that a differentiable function $g : \boldsymbol{\Theta} \to \mathbb{R}$ is L-Lipschitz smooth if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\|\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.
[12] w.r.t. the random choices $\mathbf{s}_1, \mathbf{s}_2, \dots$ of DSGD

D Supplementary Materials for Section 5

The following can be viewed as an extension of Lemma 2.3(2) and is useful for Lemma D.3:

Lemma D.1. *If $f : \boldsymbol{\Theta} \times \mathbb{S} \to \mathbb{R}$ is continuous and satisfies*

$$\sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times\mathbb{S}} \|\mathbf{z}\|^{n+3} \cdot |f(\boldsymbol{\theta},\mathbf{z})| < \infty$$

then $\int_{\mathbb{S}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |f(\boldsymbol{\theta},\mathbf{z})| \,\mathrm{d}\mathbf{z} < \infty$.

Proof. Let $g(\mathbf{z}) \coloneqq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |f(\boldsymbol{\theta},\mathbf{z})|$ and $U_k \coloneqq \mathbb{S} \cap [-k,k]^n$. Therefore,

$$\begin{aligned}
&\int_{\mathbb{S}} g(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \int_{U_1} g(\mathbf{z})\,\mathrm{d}\mathbf{z} + \int_{\mathbb{S}\setminus U_1} g(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\leq 2^n \cdot \sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times U_1} |f(\boldsymbol{\theta},\mathbf{z})| + \left(\sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times\mathbb{S}} \|\mathbf{z}\|^{n+3} \cdot f(\boldsymbol{\theta},\mathbf{z})\right) \cdot \int_{\mathbb{S}\setminus U_1} \frac{1}{\|\mathbf{z}\|^{n+3}}\,\mathrm{d}\mathbf{z}
\end{aligned}$$

All the terms are finite because (NB if $\mathbf{z} \in \mathbb{S}\setminus U_k$ then $k \leq \|\mathbf{z}\|$)

$$\begin{aligned}
\int_{\mathbb{S}\setminus U_1} \frac{1}{\|\mathbf{z}\|^{n+3}}\,\mathrm{d}\mathbf{z} &= \sum_{k=1}^{\infty} \int_{U_{k+1}\setminus U_k} \frac{1}{\|\mathbf{z}\|^{n+3}}\,\mathrm{d}\mathbf{z} \\
&\leq \sum_{k=1}^{\infty} \int_{U_{k+1}} \frac{1}{k^{n+3}}\,\mathrm{d}\mathbf{z} \\
&\leq \sum_{k=1}^{\infty} \frac{(2k)^{n+1}}{k^{n+3}}\,\mathrm{d}\mathbf{z} \\
&\leq 4^n \cdot \sum_{k=1}^{\infty} \frac{1}{k^2}\,\mathrm{d}\mathbf{z} < \infty
\end{aligned}$$

□

Lemma D.2. *If $f : \mathbb{S} \to \mathbb{R}$, where $\mathbb{S} \subseteq \mathbb{R}^n$, is a Schwartz function, $\boldsymbol{\phi}_{(-)} : \boldsymbol{\Theta}\times\mathbb{R}^n \to \mathbb{R}^n$ satisfies Assumptions 2.4 and 5.2, and $p : \mathbb{R}^n \to \mathbb{R}$ a polynomial then*

$$\int_{\mathbb{S}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) \cdot p(\mathbf{z}) \right| \mathrm{d}\mathbf{z} < \infty$$

Proof. Since $\boldsymbol{\phi}_{(-)}$ satisfies Assumption 5.2, there exists $c > 0$ such that $|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})| \leq c$ and by Assumption 2.4 and Lemma B.2 there exists a polynomial $q : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$\left| \|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})\|^{n+3} + p(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})) \right| \leq q(\mathbf{s})$$

for all $(\boldsymbol{\theta},\mathbf{s}) \in \boldsymbol{\Theta}\times\mathbb{S}$. Hence,

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times\mathbb{S}} \|\mathbf{z}\|^{n+3} \cdot \left| f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) \cdot p(\mathbf{z}) \right| \\
&= \sup_{(\boldsymbol{\theta},\mathbf{s})\in\boldsymbol{\Theta}\times\mathbb{S}} \|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})\|^{n+3} \cdot |f(\mathbf{s})| \cdot \underbrace{\left|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))\right|}_{\leq c} \cdot |p(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))| \\
&= c \cdot \sup_{\mathbf{s}\in\mathbb{S}} q(\mathbf{s}) \cdot |f(\mathbf{s})| < \infty
\end{aligned}$$

by definition of Schwartz functions and the claim follows with Lemma D.1. □

Lemma D.3. *If $f : \mathbb{S} \to \mathbb{R}$ is a Schwartz function, $\boldsymbol{\phi}_{(-)} : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies Assumptions 2.4 and 5.2, and $p : \mathbb{R}^n \to \mathbb{R}$ a polynomial then*

$$\int_{\mathbb{S}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{\partial f_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p(\mathbf{z}) \right| \mathrm{d}\mathbf{z} < \infty$$
$$\int_{\mathbb{S}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{\partial f_{(-)}}{\partial z_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p(\mathbf{z}) \right| \mathrm{d}\mathbf{z} < \infty$$
$$\int_{\mathbb{S}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{\partial^2 f_{(-)}}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}, \mathbf{z}) \cdot p(\mathbf{z}) \right| \mathrm{d}\mathbf{z} < \infty$$

where $f_{\boldsymbol{\theta}}(\mathbf{z}) \coloneqq f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|$.

Proof. As for Lemma D.2, by Lemma D.1 it suffices to prove

$$\sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times\mathbb{S}} \|\mathbf{z}\|^{n+2} \cdot \left| \frac{\partial f_{(-)}}{\partial \theta_i} \cdot p(\mathbf{z}) \right|$$

for the first claim. Note that

$$0 = \mathbf{J}_{\theta_i}(\boldsymbol{\phi}_{\boldsymbol{\theta}} \circ \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1})(\mathbf{z}) = \mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) + \mathbf{J}_{\mathbf{s}}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})$$

and hence,

$$\begin{aligned}
\mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) &= -\left(\mathbf{J}_{\mathbf{s}}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))\right)^{-1} \cdot \mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \\
&= -\frac{1}{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))|} \cdot \operatorname{adj}(\mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))) \cdot \mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \\
&= -\frac{1}{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))|} \cdot \mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))
\end{aligned}$$

for a suitable[13] function $\mathbf{g} : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}^n$ bounded by polynomials (component-wise). Therefore,

$$\begin{aligned}
\frac{\partial(f \circ \boldsymbol{\phi}_{(-)}^{-1})}{\partial \theta_i}(\mathbf{z}) &= \nabla f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \mathbf{J}_{\theta_i}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) \\
&= \frac{\nabla f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))}{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))|}
\end{aligned}$$

By Assumption 5.2, there exists $c > 0$ satisfying

$$\frac{1}{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))|} = |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})| \leq c$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{z} \in \mathbb{S}$. Consequently,

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\mathbf{z})\in\boldsymbol{\Theta}\times\mathbb{S}} \left| \|\mathbf{z}\|^{n+2} \cdot \frac{\partial(f \circ \boldsymbol{\phi}_{(-)}^{-1})}{\partial \theta_i}(\mathbf{z}) \cdot \det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) \cdot p(\mathbf{z}) \right| \\
&\leq c^2 \cdot \sup_{(\boldsymbol{\theta},\mathbf{s})\in\boldsymbol{\Theta}\times\mathbb{S}} \left| \|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})\|^{n+2} \cdot (\nabla f(\mathbf{s}) \cdot \mathbf{g}(\boldsymbol{\theta}, \mathbf{s})) \cdot p(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})) \right| < \infty
\end{aligned}$$

because $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s})$, $\mathbf{g}(\boldsymbol{\theta}, \mathbf{s})$ and $p(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))$ are uniformly bounded by polynomials independent of $\boldsymbol{\theta}$ (by Lemma B.2) and derivatives of Schwartz functions are Schwartz functions, too.

Likewise, note that

$$|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})| = \frac{1}{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))|} = \frac{1}{h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))}$$

[13] adj is the adjugate matrix

for a function $h : \boldsymbol{\Theta} \times \mathbb{R}^n \to \mathbb{R}$, the partial derivatives of which are bounded by polynomials and which we can assume by Assumption 2.4 w.l.o.g. to be positive (and greater than the constant c above).

Thus,

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) &= \frac{\partial}{\partial \theta_i} \frac{1}{h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))} \\
&= -\frac{\frac{\partial h}{\partial \theta_i}(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) + \nabla_{\mathbf{s}} h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \mathbf{J}_{\theta_i} \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})}{\left(h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))\right)^2} \\
&= -\frac{\frac{\partial h}{\partial \theta_i}(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))}{\underbrace{\left(h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))\right)^2}_{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|^{-2}}} - \frac{\nabla_{\mathbf{s}} h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))}{\underbrace{\left(h(\boldsymbol{\theta}, \boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))\right)^3}_{|\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|^{-3}}}
\end{aligned}$$

to show

$$\sup_{(\boldsymbol{\theta}, \mathbf{z}) \in \boldsymbol{\Theta} \times \mathbb{S}} \left| \|\mathbf{z}\|^{n+2} \cdot f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot \frac{\partial}{\partial \theta_i} \det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}) \cdot p(\mathbf{z}) \right| < \infty$$

The same insights can be used to show the second and third bounds. □

Corollary D.4 (Lipschitz Smoothness)**.** *If* $F \in \mathrm{Expr}$ *then the function*

$$\boldsymbol{\theta} \mapsto \mathbb{E}_{\mathbf{s} \sim \mathcal{D}}[\llbracket F \rrbracket(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]$$

is Lipschitz smooth.

In the same manner we can prove the other (simpler) obligations of (D2).

D.1 Supplementary Materials for Section 5.1

Proposition 5.4 (Uniform Convergence)**.** *If* $F \in \mathrm{SExpr}$ *then*

$$\begin{aligned}
\mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\boldsymbol{\theta}}}[\llbracket F \rrbracket_\eta(\mathbf{z})] &\xrightarrow{\text{unif.}} \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\boldsymbol{\theta}}}[\llbracket F \rrbracket(\mathbf{z})] \\
\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\boldsymbol{\theta}}}[\llbracket F \rrbracket_\eta(\mathbf{z})] &\xrightarrow{\text{unif.}} \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\boldsymbol{\theta}}}[\llbracket F \rrbracket(\mathbf{z})]
\end{aligned}$$

as $\eta \searrow 0$ *for* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*.*

Proof. Let $p : \mathbb{R}^n \to \mathbb{R}$ be a polynomial bound to $\llbracket F \rrbracket_\eta$ and $\llbracket F \rrbracket$ and let $\epsilon > 0$. We focus on the second result (the first is analogous). We define

$$\begin{aligned}
c &:= \int_{\mathbb{S}} \left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\partial \mathcal{D}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
U_k &:= \left\{ \mathbf{z} \in \mathbb{R}^n \mid |f_\eta(\mathbf{z}) - f(\mathbf{z})| > \frac{\epsilon}{2c} \text{ for some } 0 < \eta < \frac{1}{k} \right\} \\
\mu(U) &:= \int_U p(\mathbf{z}) \cdot \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial \mathcal{D}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{z}) \right| \mathrm{d}\mathbf{z}
\end{aligned}$$

which is a finite measure by Lemma D.3. Note that $(U_k)_{k \in \mathbb{N}}$ is a non-increasing sequence of sets and $\bigcap_{k \in \mathbb{N}} U_k$ is negligible. Hence, by continuity from above (of μ) there exists k such that $\mu(U_k) < \frac{\epsilon}{4}$. Finally, it suffices to observe that for $0 < \eta < \frac{1}{k}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$\begin{aligned}
&\left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))] - \frac{\partial}{\partial\theta_i}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[[\![F]\!](\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))]\right| \\
&= \left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!]_\eta(\mathbf{z})] - \frac{\partial}{\partial\theta_i}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_{\boldsymbol{\theta}}}[[\![F]\!](\mathbf{z})]\right| \\
&\le \int_{\mathbb{S}}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\cdot|[\![F]\!]_\eta(\mathbf{z}) - [\![F]\!](\mathbf{z})|\,\mathrm{d}\mathbf{z} \\
&\le \int_{U_k}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\cdot|[\![F]\!]_\eta(\mathbf{z}) - [\![F]\!](\mathbf{z})|\,\mathrm{d}\mathbf{z} \\
&\qquad + \int_{\mathbb{S}\setminus U_k}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\cdot|[\![F]\!]_\eta(\mathbf{z}) - [\![F]\!](\mathbf{z})|\,\mathrm{d}\mathbf{z} \\
&\le 2\cdot\int_{U_k}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\cdot p(\mathbf{z})\,\mathrm{d}\mathbf{z} + \int_{\mathbb{S}\setminus U_k}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\cdot\frac{\epsilon}{2c}\,\mathrm{d}\mathbf{z} \\
&\le 2\cdot\mu(U_k) + \frac{\epsilon}{2c}\cdot\int_{\mathbb{S}}\left|\frac{\partial\mathcal{D}_{(-)}}{\partial\theta_i}(\boldsymbol{\theta},\mathbf{z})\right|\mathrm{d}\mathbf{z} \\
&\le \epsilon
\end{aligned}$$

□

Uniform convergence may fail if $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is not compact:

Example D.5. Let $\boldsymbol{\Theta} := \mathbb{R}\times\mathbb{R}_{>0}$. $\mathbb{E}_{z\sim\mathcal{N}(\theta_1,\theta_2)}[\sigma_\eta(z)]$ does *not* converge uniformly to $\mathbb{E}_{z\sim\mathcal{N}(\theta_1,\theta_2)}[[z \ge 0]]$: Suppose $\eta > 0$. There exists $\delta > 0$ such that $\sigma_\eta(\delta) = 0.6$. Define $\theta_1 := \theta_2 := \frac{\delta}{2}$. Now, it suffices to note that

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{N}(\theta_1,\theta_2)}[[z \ge 0]] &\ge \int_{\theta_1-\theta_2}^{\infty}\mathcal{N}(z\mid\theta_1,\theta_2)\,\mathrm{d}z \ge 0.83 \\
\mathbb{E}_{z\sim\mathcal{N}(\theta_1,\theta_2)}[\sigma_\eta(z)] &\le 0.5\int_{-\infty}^{\theta_1-\theta_2}\mathcal{N}(z\mid\theta_1,\theta_2)\,\mathrm{d}z + 0.6\int_{\theta_1-\theta_2}^{\theta_1+\theta_2}\mathcal{N}(z\mid\theta_1,\theta_2)\,\mathrm{d}z + \int_{\theta_1+\theta_2}^{\infty}\mathcal{N}(z\mid\theta_1,\theta_2)\,\mathrm{d}z \\
&\le 0.5\cdot 0.16 + 0.6\cdot 0.69 + 0.16 = 0.654
\end{aligned}$$

D.2 Supplementary Materials for Section 5.2

Now, exploiting the chain rule and Assumption 2.4, it is straightforward to show inductively that for $F \in \mathrm{Expr}_\ell$,

Lemma D.6. *If $F \in \mathrm{Expr}_\ell$ there exists a polynomial $p : \mathbb{R}^n \to \mathbb{R}$ such that for all $\eta > 0$ and $\mathbf{z} \in \mathbb{R}^n$,*

$$\left|\frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right| \le |\sigma'_\eta|^\ell \cdot p(\mathbf{z})$$

(By $|\sigma'_\eta|$ we mean $\sup_{y\in\mathbb{R}}|\sigma'_\eta(y)|$.)

Lemma D.7. *Let $F \in \mathrm{Expr}_\ell$, $f : \mathbb{S} \to \mathbb{R}$ be a non-negative Schwartz function, where $\mathbb{S} = U_1 \times \cdots \times U_n$ and $U_1, \dots, U_n \in \{\mathbb{R}, \mathbb{R}_{\ge 0}\}$, and p be a non-negative polynomial. For all $1 \le i \le n$ there exists $c > 0$ such that for all $0 < \eta \le 1$,*

$$\begin{aligned}
\int_{\mathbb{S}} f_{\boldsymbol{\theta}}(\mathbf{z})\cdot p(\mathbf{z})\cdot\left|\frac{\partial^2[\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z})\right|\mathrm{d}\mathbf{z} &< c\cdot\eta^{-\ell} \\
\int_{\mathbb{S}} f_{\boldsymbol{\theta}}(\mathbf{z})\cdot p(\mathbf{z})\cdot\left|\frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right|^2\mathrm{d}\mathbf{z} &< c\cdot\eta^{-\ell}
\end{aligned}$$

where $f_{\boldsymbol{\theta}} := f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}))\cdot|\det\mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|$.

Proof. Note that $f_{\boldsymbol{\theta}}$ is differentiable and non-negative. To simplify notation, we assume that $\mathbb{S} = \mathbb{R}^n$. (Otherwise the proof is similar, exploiting Lemma D.6.) Besides, it suffices to establish the first bound. To see that the second is a consequence of the first, we use integration by parts

$$\begin{aligned}
&\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left(\frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right)^2 \mathrm{d}\mathbf{z} \\
&= \int \int \left(f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \,\mathrm{d}z_j \,\mathrm{d}\mathbf{z}_{-j} \\
&= \int \underbrace{\left[f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot [\![F]\!]_\eta(\mathbf{z})\right]_{-\infty}^{\infty}}_{=0} \mathrm{d}\mathbf{z}_{-j} \\
&\qquad - \int \frac{\partial}{\partial z_i}\left(f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right) \cdot [\![F]\!]_\eta(\mathbf{z}) \,\mathrm{d}\mathbf{z}
\end{aligned}$$

where $\mathbf{z}_{-j}$ is $z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, j_n$, because by Lemma D.2, for fixed $\eta > 0$ and $\mathbf{z}_{-j}$,

$$\lim_{z_j \to \pm\infty} f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \underbrace{p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot [\![F]\!]_\eta(\mathbf{z})}_{\text{bounded by polynomial}}$$

We continue bounding:

$$\begin{aligned}
&- \int \frac{\partial}{\partial z_i}\left(f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z})\right) \cdot [\![F]\!]_\eta(\mathbf{z}) \,\mathrm{d}\mathbf{z} \\
&= - \int \frac{\partial f_{(-)}}{\partial z_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot [\![F]\!]_\eta(\mathbf{z}) \,\mathrm{d}\mathbf{z} \\
&\qquad - \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \frac{\partial p}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial[\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot [\![F]\!]_\eta(\mathbf{z}) \,\mathrm{d}\mathbf{z} \\
&\qquad - \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \frac{\partial^2[\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z}) \cdot [\![F]\!]_\eta(\mathbf{z}) \,\mathrm{d}\mathbf{z} \\
&\leq \eta^{-\ell} \cdot \int \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\frac{\partial f_{(-)}}{\partial z_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p_1(\mathbf{z})\right| \mathrm{d}\mathbf{z} \\
&\qquad + \eta^{-\ell} \cdot \int \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z})| \,\mathrm{d}\mathbf{z} \\
&\qquad + \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_3(\mathbf{z}) \cdot \left|\frac{\partial^2[\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z})\right|
\end{aligned}$$

for suitable polynomial bounds $p_1, p_2, p_3 : \mathbb{R}^n \to \mathbb{R}$ (which exist due to the fact that $\frac{\partial p}{\partial z_i}$ *is* a polynomial, Item 1 and Lemma D.6) and the second inequality follows with Lemma D.3.

We prove the claim by induction on the definition of Expr_ℓ:

- For $z_j \in \mathrm{Expr}_0$ and $F \in \mathrm{Expr}_{\ell+1}$ due to $F \in \mathrm{Expr}_\ell$ the claim is obvious.
- For $g(F_1, \ldots, F_k) \in \mathrm{Expr}_\ell$ because $F_1, \ldots, f_k \in \mathrm{Expr}_\ell$,

$$\begin{aligned}
&\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left|\frac{\partial^2(g \circ \langle [\![F_1]\!]_\eta, \ldots, [\![F_k]\!]_\eta \rangle)}{\partial z_i^2}(\mathbf{z})\right| \mathrm{d}\mathbf{z} \\
&\leq \sum_{1 \leq j, j' \leq k} \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left|\frac{\partial^2 g}{\partial y_j \partial y_{j'}}([\![F_1]\!]_\eta(\mathbf{z}), \ldots, [\![F_k]\!]_\eta(\mathbf{z})) \cdot \frac{\partial[\![F_j]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial[\![F_{j'}]\!]_\eta}{\partial z_i}(\mathbf{z})\right| \mathrm{d}\mathbf{z} \\
&\quad + \sum_{1 \leq j \leq k} \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left|\frac{\partial g}{\partial y_j}([\![F_1]\!]_\eta(\mathbf{z}), \ldots, [\![F_k]\!]_\eta(\mathbf{z})) \cdot \frac{\partial^2[\![F_j]\!]_\eta}{\partial^2 z_i}(\mathbf{z})\right| \mathrm{d}\mathbf{z}
\end{aligned}$$

  By Assumption 2.4, $\frac{\partial g}{\partial y_j}(g_\eta^{(1)}(\mathbf{z}), \ldots, g_\eta^{(\ell)}(\mathbf{z}))$ is bounded by a polynomial. Therefore the second summand can be bounded by the inductive hypothesis. For the first, again by Assumption 2.4, we can bound

$$p(\mathbf{z}) \cdot \left|\frac{\partial^2 g}{\partial y_j \partial y_{j'}}(g_\eta^{(1)}(\mathbf{z}), \ldots, g_\eta^{(\ell)}(\mathbf{z}))\right| \leq p_1(\mathbf{z})$$

for a (non-negative) polynomial p_1, apply the Cauchy-Schwarz inequality and apply the inductive hypothesis to

$$\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_1(\mathbf{z}) \cdot \left| \frac{\partial [\![F_j]\!]_\eta}{\partial z_i}(\mathbf{z}) \right|^2 \mathrm{d}\mathbf{z} \qquad \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_1(\mathbf{z}) \cdot \left| \frac{\partial [\![F_{j'}]\!]_\eta}{\partial z_i}(\mathbf{z}) \right|^2 \mathrm{d}\mathbf{z}$$

- Next, suppose $\mathbf{if}\, F_1 < 0 \,\mathbf{then}\, F_2 \,\mathbf{else}\, F_3 \in \mathrm{Expr}_{\ell+1}$ because $F_1 \in \mathrm{Expr}_\ell$ and $F_2, F_3 \in \mathrm{Expr}_{\ell+1}$. By linearity we bound (similarly for the other branch):

$$\begin{aligned}
&\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left| \frac{\partial^2((\sigma_\eta \circ [\![F_1]\!]_\eta) \cdot [\![F_3]\!]_\eta)}{\partial z_i^2}(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\le \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left| \frac{\partial^2(\sigma_\eta \circ [\![F_1]\!]_\eta)}{\partial z_i^2}(\mathbf{z}) \cdot [\![F_3]\!]_\eta(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\quad + 2 \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left| \frac{\partial(\sigma_\eta \circ [\![F_1]\!]_\eta)}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial [\![F_3]\!]_\eta}{\partial z_i}(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\quad + \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left| \sigma_\eta([\![F_1]\!]_\eta(\mathbf{z})) \cdot \frac{\partial^2 [\![F_3]\!]_\eta}{\partial z_i^2}(\mathbf{z}) \right| \mathrm{d}\mathbf{z}
\end{aligned}$$

We abbreviate $F \equiv F_1$ and bound $p(\mathbf{z}) \cdot |[\![F_3]\!]_\eta(\mathbf{z})|$ by the non-negative polynomial p_2. Bounding the first summand is most interesting:

$$\begin{aligned}
&\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p(\mathbf{z}) \cdot \left| \frac{\partial^2(\sigma_\eta \circ [\![F_1]\!]_\eta)}{\partial z_i^2}(\mathbf{z}) \cdot [\![F_3]\!]_\eta(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&= \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \left| \frac{\partial^2(\sigma_\eta \circ [\![F]\!]_\eta)}{\partial z_i}(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\le \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \left| \sigma_\eta''([\![F]\!]_\eta(\mathbf{z})) \right| \cdot \left(\frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \right)^2 \mathrm{d}\mathbf{z} \\
&\quad + \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \left| \underbrace{\sigma_\eta'([\![F]\!]_\eta(\mathbf{z}))}_{\le \eta^{-1}} \cdot \frac{\partial^2 [\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z}) \right| \mathrm{d}\mathbf{z}
\end{aligned}$$

The first summand can be bounded with the inductive hypothesis. For the second summand we exploit that

$$\begin{aligned}
|\sigma_\eta''(x)| &= \eta^{-2} \left| \sigma_\eta(x) \cdot (1 - \sigma_\eta(x)) \cdot (1 - 2\sigma_\eta(x)) \right| \\
&\le \eta^{-2} \cdot \sigma_\eta(x) \cdot (1 - \sigma_\eta(x)) \\
&= \eta^{-1} \cdot \sigma_\eta'(x)
\end{aligned}$$

and continue using integration by parts again in the second step

$$
\begin{aligned}
&\int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \left| \sigma''_\eta([\![F]\!]_\eta(\mathbf{z})) \cdot \left(\frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \right)^2 \right| \mathrm{d}\mathbf{z} \\
&\leq \eta^{-1} \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial (\sigma_\eta \circ [\![F]\!]_\eta)}{\partial z_i}(\mathbf{z}) \,\mathrm{d}\mathbf{z} \\
&= -\eta^{-1} \cdot \int \frac{\partial}{\partial z_i} \left(f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \right) \cdot \sigma_\eta([\![F]\!]_\eta(\mathbf{z})) \,\mathrm{d}\mathbf{z} \\
&= -\eta^{-1} \cdot \int \frac{\partial f_{(-)}}{\partial z_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot \sigma_\eta([\![F]\!]_\eta(\mathbf{z})) \,\mathrm{d}\mathbf{z} \\
&\qquad - \eta^{-1} \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \frac{\partial p_2}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \cdot \sigma_\eta([\![F]\!]_\eta(\mathbf{z})) \,\mathrm{d}\mathbf{z} \\
&\qquad - \eta^{-1} \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \frac{\partial^2 [\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z}) \cdot \sigma_\eta([\![F]\!]_\eta(\mathbf{z})) \,\mathrm{d}\mathbf{z} \\
&\leq \eta^{-1} \cdot \int \left| \frac{\partial f_{(-)}}{\partial z_i}(\boldsymbol{\theta}, \mathbf{z}) \cdot p_2(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\qquad + \eta^{-1} \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \left| \frac{\partial p_2}{\partial z_i}(\mathbf{z}) \cdot \frac{\partial [\![F]\!]_\eta}{\partial z_i}(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \\
&\qquad + \eta^{-1} \cdot \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \left| p_2(\mathbf{z}) \cdot \frac{\partial^2 [\![F]\!]_\eta}{\partial z_i^2}(\mathbf{z}) \right| \mathrm{d}\mathbf{z}
\end{aligned}
$$

and the claim follows with the inductive hypothesis (recall $F \equiv F_1 \in \mathrm{Expr}_\ell$), Lemmas D.3 and D.6. ☐

Proposition 5.5. *If $F \in \mathrm{Expr}_\ell$ then there exists $L > 0$ such that for all $\eta > 0$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}[\|\nabla_{\boldsymbol{\theta}} [\![F]\!]_\eta(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{s}))\|^2] \leq L \cdot \eta^{-\ell}$.*

Proof. Note that

$$
\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[\left| \frac{\partial ([\![F]\!]_\eta \circ \boldsymbol{\phi}_{(-)})}{\partial \theta_i}(\mathbf{s}) \right|^2 \right] = \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[\left| \sum_{j=1}^n \frac{\partial [\![F]\!]_\eta}{\partial z_j}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \frac{\partial \phi^{(j)}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{s}) \right|^2 \right]
$$

Therefore, by the Cauchy-Schwarz inequality it suffices to bound

$$
\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[\left| \frac{\partial [\![F]\!]_\eta}{\partial z_j}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \frac{\partial \phi^{(j)}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{s}) \right|^2 \right]
$$

for all $1 \leq j \leq n$. By Lemma B.2, for each $1 \leq i \leq m$ and $1 \leq j \leq n$ there exists a (non-negative) polynomial bound $p : \mathbb{R}^n \to \mathbb{R}$ satisfying

$$
\left(\frac{\partial \phi^{(j)}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{s}) \right)^2 \leq p(\mathbf{s})
$$

and for $f(\mathbf{s}) \coloneqq \mathcal{D}(\mathbf{s}) \cdot p(\mathbf{s})$, which is a Schwartz function by Item 3,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[\left| \frac{\partial [\![F]\!]_\eta}{\partial z_j}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \cdot \frac{\partial \phi^{(j)}_{(-)}}{\partial \theta_i}(\boldsymbol{\theta}, \mathbf{s}) \right|^2 \right] &\leq \int f(\mathbf{s}) \cdot \left| \frac{\partial [\![F]\!]_\eta}{\partial z_j}(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{z})) \right|^2 \mathrm{d}\mathbf{s} \\
&= \int f_{\boldsymbol{\theta}}(\mathbf{z}) \cdot \left| \frac{\partial [\![F]\!]_\eta}{\partial z_j}(\mathbf{z}) \right|^2 \mathrm{d}\mathbf{z}
\end{aligned}
$$

where

$$
f_{\boldsymbol{\theta}} \coloneqq f(\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) \cdot |\det \mathbf{J}\boldsymbol{\phi}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})|
$$

The claim follows with Lemma D.7. ☐

D.3 Supplementary Materials for Section 5.3

Remark D.8 (Average Variance of Run). By Proposition 5.5 the average variance of a finite DSGD run with length N is bounded by (using Hölder's inequality)

$$\frac{1}{N} \cdot \sum_{k=1}^{N} \eta_k^{-\ell} \leq \eta_{\frac{N+1}{2}}^{-\ell}$$

Consequently, the average variance of a DSGD run is lower than for standard SGD with a fixed accuracy coefficient $\eta < \eta_{\frac{N+1}{2}}$, the accuracy coefficient of DSGD after half the iterations.

Proof. If $\eta_k \in \Theta(k^{-\frac{1}{\ell}+\epsilon})$ then (modulo constants), $\eta_k^{-\ell} = k^\delta$ for $\delta \leq 1$. Therefore, by Hölder's inequality,

$$\begin{aligned}
\frac{1}{N} \cdot \sum_{k=1}^{N} \eta_k^{-\ell} &= \frac{1}{N} \cdot \sum_{k=1}^{N} k^\delta \\
&\leq \frac{1}{N} \cdot \left(N^{1/\delta - 1} \cdot \sum_{k=1}^{N} k \right)^\delta \\
&= \frac{1}{N} \cdot \left(N^{1/\delta - 1} \cdot \frac{N \cdot (N+1)}{2} \right)^\delta \\
&= \left(\frac{N+1}{2} \right)^\delta = \eta_{\frac{N+1}{2}}^{-\ell}
\end{aligned}$$

□

E Supplementary Materials for Section 6

The code is available at `https://github.com/domwagner/DSGD.git`. The experiements can be viewed and run in the jupyter notebook `experiments.ipynb` by running:

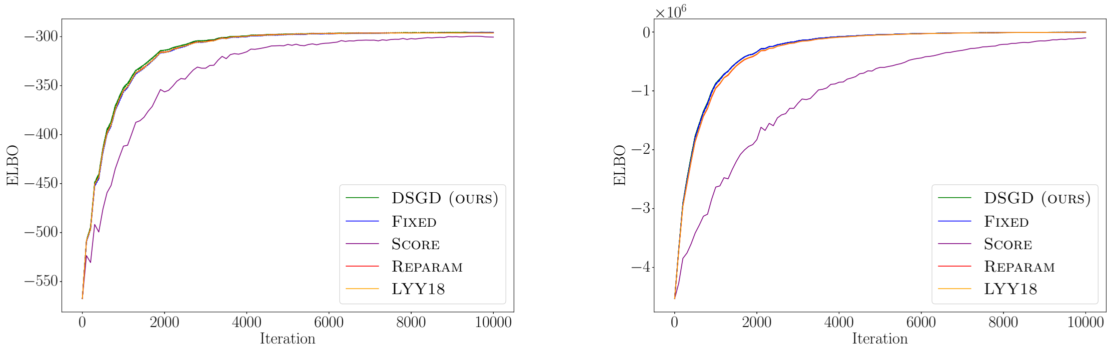
`jupyter notebook experiments.ipynb`.

E.1 Additional Models

- `cheating` (Davidson-Pilon, 2015) simulates a differential privacy setting where students taking an exam are surveyed to determine the prevalence of cheating without exposing the details for any individual. Students are tasked to toss a coin, on heads they tell the truth (cheating or not cheating) and on tails they toss a second coin to determine their answer. The tossing of coins here is a source of discontinuity. The goal, given the proportion of students who answered yes, is to predict a posterior on the cheating rate. In this model there are 300 if-statements and a 301-dimensional latent space, although we only optimise over a single dimension with the other 300 being sources of randomness.
- `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 (the two rates and the point at which they change) and 37 if-statements.
- `influenza` (Shumway and Stoffer, 2005) models the US influenza mortality for 1969. In each month, the mortality rate depends on the dominant virus strain being of type 1 or type 2, producing a non-differentiablity for each month. Given the mortality data, the goal is to infer the dominant virus strain in each month. The model has a 37-dimensional latent variable and 24 if-statements.

E.2 Additional Results

See Fig. 3 and Tables 4 and 5.



(a) `textmsg` (b) `influenza`

Figure 3: ELBO trajectories for additional models. The same setup is used as in Fig. 2.

Table 4: Computational cost and work-normalised variances for additional models (cf. Table 2)

(a) `textmsg`

Estimator	Cost	Avg($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	1.84	7.89e-03	1.53e-02
Fixed	1.79	1.08e-02	2.14e-02
Reparam	1.25	7.99e-03	1.53e-02
LYY18	4.51	3.42e-02	6.00e-02

(b) `influenza`

Estimator	Cost	Avg($V(.)$)	$V(\|.\|_2)$
DSGD (ours)	1.28	7.77e-03	3.94e-03
Fixed	1.28	7.92e-03	3.97e-03
Reparam	1.21	7.60e-03	3.75e-03
LYY18	8.30	5.80e-02	2.88e-02

Table 5: Mean of the final ELBO (the higher the better) for different random seeds and indicating error bars (the ± is one standard deviation).

(a) `temperature`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM	LYY18
0.06	-76 ± 1	-624,250 ± 44,121			
0.1	-84 ± 2	-425 ± 9			
0.14	-15,476 ± 4,641	-121,932 ± 85,460	-2,611,479 ± 255,193	-706,729 ± 4,697	-17,502 ± 52,044
0.18	-94,125 ± 6,930	-32,171 ± 66			
0.22	-165,787 ± 9,758	-155,321 ± 61,732			

(b) `xornet`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM
0.06	-3,530 ± 3,889	-5,522 ± 4,136		
0.1	-33 ± 7	-2,029 ± 3,305		
0.14	-27 ± 4	-2,028 ± 3,986	-553 ± 1,507	-9,984 ± 38
0.18	-25 ± 3	-26 ± 6		
0.22	-30 ± 8	-26 ± 2		

(c) `random-walk`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM	LYY18
0.06	-37 ± 148	-85 ± 197			
0.1	-37 ± 148	-86 ± 197			
0.14	-38 ± 148	-37 ± 148	-85 ± 197	-371,612 ± 8,858	-135 ± 226
0.18	-38 ± 148	-37 ± 148			
0.22	-38 ± 148	-37 ± 148			

(d) `cheating`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM	LYY18
0.06	-65 ± 1	-65 ± 1			
0.1	-65 ± 1	-65 ± 1			
0.14	-65 ± 1	-65 ± 1	-66 ± 1	-80 ± 1	-65 ± 1
0.18	-65 ± 1	-65 ± 1			
0.22	-65 ± 1	-65 ± 1			

(e) `textmsg`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM	LYY18
0.06	-295 ± 1	-295 ± 1			
0.1	-295 ± 1	-296 ± 1			
0.14	-295 ± 1	-296 ± 1	-300 ± 1	-296 ± 1	-296 ± 1
0.18	-296 ± 1	-296 ± 1			
0.22	-296 ± 1	-296 ± 1			

(f) `influenza`

η/η_{4000}	DSGD (ours)	FIXED	SCORE	REPARAM	LYY18
0.06	-3,586 ± 112	-3,589 ± 114			
0.1	-3,584 ± 111	-3,590 ± 111			
0.14	-3,582 ± 111	-3,590 ± 111	-95,380 ± 3,567	-4,045 ± 108	-3,860 ± 106
0.18	-3,579 ± 111	-3,589 ± 111			
0.22	-3,577 ± 111	-3,587 ± 112			