A Language and Smoothed Semantics for Convergent Stochastic Gradient Descent

Dominik Wagner Basim Khajwal Luke Ong

UNIVERSITY OF OXFORD

Logic of Probabilistic Programming
31 January 2022

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

no dependence on $\boldsymbol{\theta}$

expressed in PL with conditionals,
may *not* be *differentiable/continuous*

Example: maximisation of ELBO for reparametrised models in *variational inference*

$$\mathrm{ELBO}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{z}\sim q}\left[\log p(\phi_{\boldsymbol{\theta}}(\mathbf{z})) - \log q_{\theta}(\phi_{\boldsymbol{\theta}}(\mathbf{z}))\right]$$

model

guide

Benefit of reparametrisation: lower variance

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

no dependence on $\boldsymbol{\theta}$

expressed in PL with conditionals,
may *not* be *differentiable/continuous*

Aim: find *stationary* point, i.e. $\boldsymbol{\theta}$ s.t. $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}\sim q}[f(\boldsymbol{\theta}, \mathbf{z})] = 0$

Stochastic Gradient Descent (SGD)

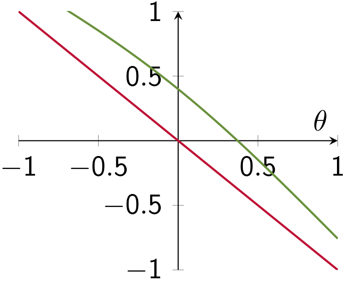
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \cdot \underbrace{\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \mathbf{z}_k)}_{\textit{reparametrisation gradient estimator}} \qquad \mathbf{z}_k \sim q$$

Reparametrisation gradient estimator for non-differentiable models is biased!

[Lee et al., NeurIPS 2018]

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

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



Vanishing gradient estimator does not imply stationarity!

Contributions

Provable convergence **to stationary points (and unbiased gradient estimators) for *typable* programs.**

Approach:

- *Smoothen* (discontinuous) function using sigmoid with accuracy coefficient
- Optimise expectation, enhancing accuracy in each step

This talk:

- *Reparametrisation* programming language
- Type system and smoothed semantics
- *Convergence* of *Diagonalisation* Stochastic Gradient Descent, a new variant of SGD
- Empirical evaluation

Part I:
Programming Language, Type System and Smoothed Semantics

Reparametrisation Programming Language

simply typed λ-calculus with $\mathbb{R}$, $+$, $\cdot$ and *conditionals*

$+$ *sampling* from standard normal
transformed by *diffeomorphic* polynomials

$$
\begin{aligned}
M ::= \quad & \cdots \\
& \mid \textbf{if } M < 0 \textbf{ then } M \textbf{ else } M \\
& \mid \ \underline{\phi}_{\boldsymbol{\theta}}(M, \dots, M, \textbf{sample})
\end{aligned}
$$

diffeomorphic polynomial

Example: sample from $\mathcal{N}(\mu, \sigma)$ using $\phi_{\mu,\sigma}(\textbf{sample})$, where $\phi_{\mu,\sigma}(z) := \sigma \cdot z + \mu$

$$\mathrm{argmin}_{\boldsymbol{\theta}}\, \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[[\![\mathbf{M}]\!](\boldsymbol{\theta},\mathbf{z})\right]$$

where $[\![M]\!]$ is the *value*-function of a term $M : R$ with parameters $\boldsymbol{\theta} : R$.

(Integrability) $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[|\,[\![M]\!]\,(\boldsymbol{\theta},\mathbf{z})|] < \infty$ for all $\boldsymbol{\theta} \in \mathbb{R}^n$.

Type System

ensure guards do not directly depend on parameters

(only after transformation)

$$\textbf{if}\ \theta < 0\ \textbf{then}\ 0\ \textbf{else}\ 1 \qquad ✗$$

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

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

$$(\lambda x.\, \underline{-0.5} \cdot \theta^2 + (\textbf{if}\ x < 0\ \textbf{then}\ \underline{0}\ \textbf{else}\ \underline{1}))\ (\underline{\phi}_\theta(\textbf{sample})) \qquad ✓$$

Two kinds of typing judgements:

$$\Gamma \vdash M : \tau$$

no parameters θ_i

$$\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau$$

use unrestricted

not in guards

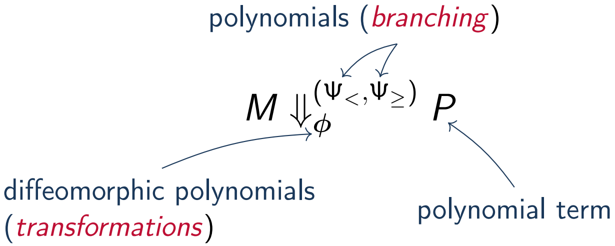
$$\frac{\Gamma \vdash L : R \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} M : \tau \quad \Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} N : \tau}{\Gamma \mid \Delta \vdash_{\boldsymbol{\theta}} \mathbf{if}\ L < 0\ \mathbf{then}\ M\ \mathbf{else}\ N : \tau}$$

Reparametrisation-aware symbolic execution

variant of [Mak et al., ESOP 2021]

- Collect constraints due to *branching*
- Replace $\underline{\phi}_{\boldsymbol{\theta}}(P_1, \ldots, P_\ell, \textbf{sample})$ with fresh sampling variable α_j and keep track of *transformations*

$\emptyset \mid \emptyset \vdash_{\boldsymbol{\theta}} M : R$

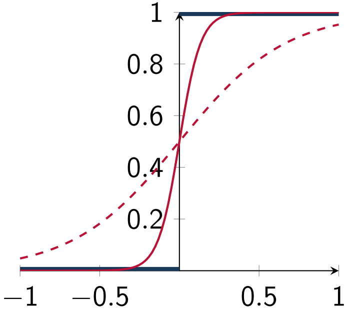


$$M \Downarrow_{\boldsymbol{\phi}}^{(\Psi_<,\Psi_\geq)} P$$

polynomials (*branching*)
diffeomorphic polynomials (*transformations*)
polynomial term

“Standard” Semantics for accuracy coefficient $k \in \mathbb{N}$

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



Smoothed Semantics for accuracy coefficient $k \in \mathbb{N}$

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

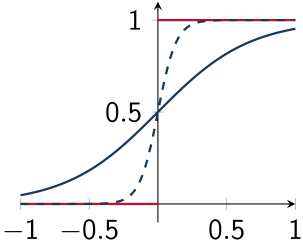
Adapt (backward mode) automatic differentiation to compute smoothing

Part II:
Properties of Smoothing

(Unbiasedness) $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}}[[\![M]\!]_k (\boldsymbol{\theta}, \mathbf{z})] = \mathbb{E}_{\mathbf{z}}[\nabla_{\boldsymbol{\theta}} [\![M]\!]_k (\boldsymbol{\theta}, \mathbf{z})]$ for all $k \in \mathbb{N}$.

Use SGD for $[\![M]\!]_k$ for fixed $k \in \mathbb{N}$

Are stationary points of $\mathbb{E}[[\![M]\!]_k (\boldsymbol{\theta}, \mathbf{z})]$ approximately stationary for $\mathbb{E}[[\![M]\!] (\boldsymbol{\theta}, \mathbf{z})]$?



$[\![M]\!]_k \to [\![M]\!]$ pointwisely as $k \to \infty$ (*not uniformly!*)

However, set of approximate roots of polynomials is "small".

(Uniform Convergence) If $\boldsymbol{\Theta} \subseteq \mathbb{R}^n$ is compact then

$$\mathbb{E}_{\mathbf{z}}[[\![\mathbf{M}]\!]_{\mathbf{k}}(\boldsymbol{\theta},\mathbf{z})] \xrightarrow{\text{unif}} \mathbb{E}_{\mathbf{z}}[[\![\mathbf{M}]\!](\boldsymbol{\theta},\mathbf{z})] \qquad \text{as } k \to \infty \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

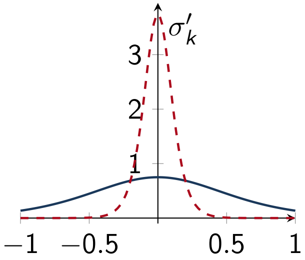
$\phi_\theta(z) := c \cdot z + \theta$, where $0 \neq c \in \mathbb{R}$

$$M \equiv \textbf{if } \underline{\phi}_\theta(\textbf{sample}) < 0 \textbf{ then } \underline{0} \textbf{ else } \underline{1}$$

$$[\![M]\!]_k\,(\theta, z) = \sigma_k(\phi_\theta(z))$$

Apply the chain rule:

$$\nabla_\theta\,[\![M]\!]_k\,(\theta, z) = \sigma'_{\mathbf{k}}(\phi_\theta(\mathbf{z}))$$



$\nabla_\theta\,[\![M]\!]_k\,(\theta, z)$ is *unbounded* whenever $\phi_\theta(z) = 0$!

$\phi_\theta(z) := c \cdot z + \theta$, where $0 \neq c \in \mathbb{R}$

$$M \equiv \textbf{if } \underline{\phi}_\theta(\textbf{sample}) < 0 \textbf{ then } \underline{0} \textbf{ else } \underline{1}$$

$$[\![M]\!]_k(\theta, z) = \sigma_k(\phi_\theta(z))$$

Apply the chain rule:

$$\nabla_\theta [\![M]\!]_k(\theta, z) = \sigma'_{\mathbf{k}}(\phi_\theta(\mathbf{z})) = \frac{1}{c} \cdot \nabla_z(\sigma_k \circ \phi_{(-)})(\theta, z)$$

Enables integration by part:

$$\begin{aligned}\mathbb{E}_z\left[\nabla_\theta [\![M]\!]_k(\theta, z)\right] &= \int \mathcal{N}(z) \cdot \frac{1}{c} \cdot \nabla_{\mathbf{z}}(\sigma_k \circ \phi_{(-)})(\theta, z)\,\mathrm{d}z \\ &= \frac{1}{c}\left(\underbrace{[\mathcal{N}(z) \cdot \sigma_k(\phi_\theta(z))]_{-\infty}^{\infty}}_{0} + \underbrace{\mathbb{E}_z[z \cdot \sigma_k(\phi_\theta(z))]}_{\xrightarrow{\text{unif}} \mathbb{E}[z \cdot [\phi_\theta(z) > 0]]}\right)\end{aligned}$$

(Uniform Convergence of Gradients) If $\boldsymbol{\Theta} \subseteq \mathbb{R}^n$ is compact then

$$\nabla_{\mathbf{z}} \mathbb{E}_{\mathbf{z}}[[\![\mathsf{M}]\!]_{\mathbf{k}}(\boldsymbol{\theta}, \mathbf{z})] \xrightarrow{\text{unif}} \nabla_{\mathbf{z}} \mathbb{E}_{\mathbf{z}}[[\![\mathsf{M}]\!](\boldsymbol{\theta}, \mathbf{z})] \qquad \text{as } k \to \infty \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

Basis for finding approximately stationary points:

For $\epsilon > 0$ exists $k \in \mathbb{N}$ s.t. stationary points $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ of the *k-smoothed* problem satisfy

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

Part III:

Diagonalisation Stochastic Gradient Descent

Diagonalisation Stochastic Gradient Descent (DSGD)

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \cdot \nabla_{\boldsymbol{\theta}} [\![\mathsf{M}]\!]_{\mathsf{k}}(\boldsymbol{\theta}_k, \mathbf{z}_k) \qquad \mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

As a consequence of unbiasedness, uniform convergence (of gradients), etc.

Convergence on Typable Programs

If $\emptyset \mid \emptyset \vdash_{\boldsymbol{\theta}} M : R$ then a DSGD sequence $(\boldsymbol{\theta}_k)_{k \in \mathbb{N}}$

1. is unbounded or
2. has a *stationary* accumulation point.

Part IV:

Evaluation

Related Work

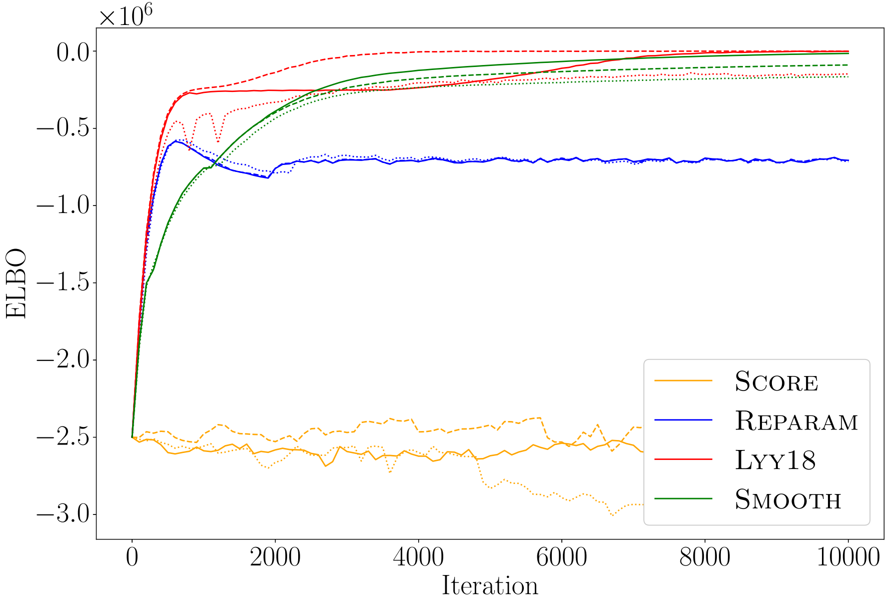
[Lee et al., NeurIPS 2018]:

- Fix (biased) reparametrisation gradient estimator for non-differentiable models by additional non-trivial *boundary* terms
- ✗ Only discuss efficient method for *affine* guards
- ✗ Not concerned with *convergence* of SGD
- ✗ No discussion of PL aspects

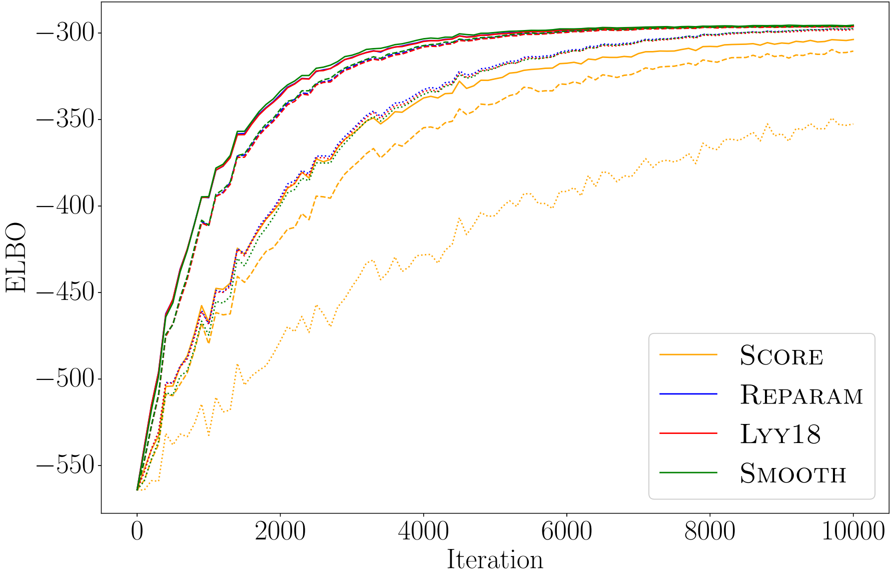
Our work:

- ✓ Type system enforcing very mild restrictions on PL
- ✓ *Simple:* smoothed semantics avoids boundary term
- ✓ Not only unbiasedness but also *convergence* of DSGD
- *Asymptotic* result, for each fixed accuracy smoothing (only) approximation

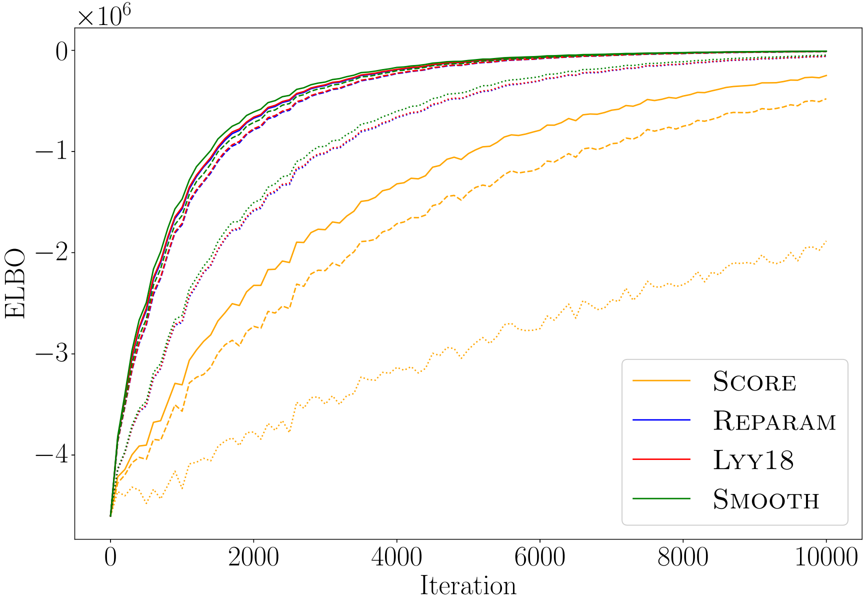
Experimental Evaluation: `temperature`



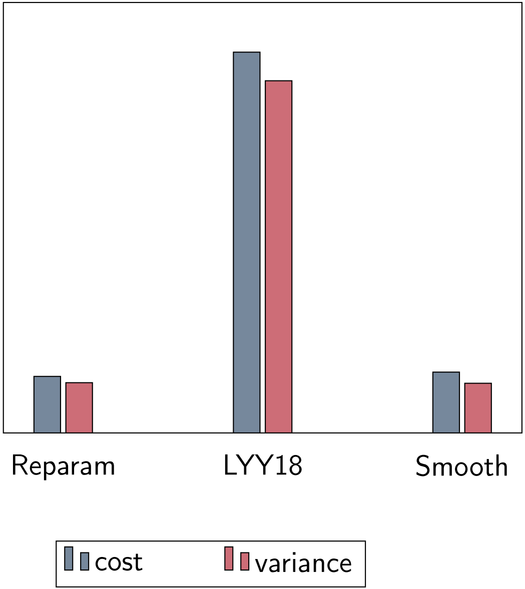
Experimental Evaluation: `textmsg`



Experimental Evaluation: `influenza`



Computational Cost and Variance: `influenza`



Conclusion

Provable convergence of Diagonalisation Stochastic Gradient Descent

- Smoothed Semantics
- Type system enforcing very mild restrictions on PL
- Unbiased gradient estimators
- Competitive on benchmarks

Future work:

- Beyond normal distributions and polynomials
- Recursion

$$\tau ::= R \mid \tau \to \tau \mid \tau_{\boldsymbol{\theta}} \to \tau$$

may depend on parameters

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

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