UNIVERSITY OF OXFORD

# A Language and Smoothed Semantics for Convergent Stochastic Gradient Descent

Dominik Wagner and Luke Ong

**Idealised Programming Language:**

$M ::= x \mid \underline{r} \mid M + M \mid M \cdot M \mid \textbf{if } M < 0 \textbf{ then } M \textbf{ else } M \mid \lambda y.\, M \mid M\, M$

## Problem Statement

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

discontinuous

where $M$ is a term of type $R$ and has free variables $\theta$ and $\mathbf{z}$ of type $R$, and $\mathcal{N}(\mathbf{0},\mathbf{I})$ is the multivariate standard normal distribution.

*Find stationary points*

**Example:** maximisation of ELBO where the variation distribution is represented as a parameterised transformation of a noise distribution

$$E_{\mathbf{z}\sim q(z)}\left[\log(p(\phi_\theta(z))) - \log q_\theta(\phi_\theta(z))\right]$$

polynomials after simplification

## Reparametrisation Gradient is Biased [LYY18]

Consider

$$M \equiv \underline{-0.5}\cdot(z+\theta)^2 + (\textbf{if } z+\theta < 0 \textbf{ then } \underline{0} \textbf{ else } \underline{1}) + \underline{0.5}\cdot z^2$$

Then

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

*Vanishing gradient estimator does not imply stationarity!*

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

**Approach:** First *smoothen* function using sigmoid with accuracy coefficient k; then optimise expectation, enhancing accuracy in each step.

$$\underline{-0.5}\cdot(z+\theta)^2 + \sigma_k(z+\theta) + \underline{0.5}\cdot z^2$$

where

$$\sigma_k(x) := \sigma\left(\frac{x}{\sqrt{k}}\right) = \frac{1}{1+\exp\left(-\frac{x}{\sqrt{k}}\right)}$$

*Keep track of branching behaviour (for definition of smoothing) and reparametrisations (for convergence proof).*

**Type system** enforces two restrictions:

1. in each branch, each $z_i$ occurs at most *once* and its *transformation* is *affine*
2. *guards* of conditionals do not contain parameters $\theta$ or (untransformed) $\mathbf{z}$

**Example:** The running example can be rephrased as $N \equiv$

$$(\lambda y.\, \underline{-0.5}\cdot y^2 + (\textbf{if } y < 0 \textbf{ then } \underline{0} \textbf{ else } \underline{1}) + \underline{0.5}\cdot(y-\theta)^2)\ \underbrace{((\lambda z.\, z+\theta)\, z)}_{\text{affine reparametrisation}}$$

## Symbolic Operational Semantics [MOPW21]

$$M \Downarrow_\phi^{\Psi_<,\Psi_\geq} \mathcal{V}$$

polynomials

affine

iff for $\theta, \mathbf{z}$ such that for $\psi\in\Psi_<$, $\psi(\phi_\theta(\mathbf{z})) < 0$, and for $\psi\in\Psi_\geq$, $\psi(\phi_\theta(\mathbf{z})) \geq 0$, it holds $[\![M]\!]\,(\theta,\mathbf{z}) = [\![\mathcal{V}]\!]\,(\theta,\mathbf{z})$.

*Sound and complete view of branching behaviour*

**Example:**

$$N \Downarrow_{z\mapsto z+\theta}^{\{y\},\emptyset} \underline{-0.5}\cdot y^2 + \underline{0} + \underline{0.5}\cdot(y-\theta)^2$$

$$N \Downarrow_{z\mapsto z+\theta}^{\emptyset,\{y\}} \underline{-0.5}\cdot y^2 + \underline{1} + \underline{0.5}\cdot(y-\theta)^2$$

## Smoothed Semantics

For *accuracy coefficient* $k\in\mathbb{N}$,

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

*Adapt (backward mode) automatic differentiation to compute smoothing*

## Diagonalisation Gradient Descent

Suppose for each $k\in\mathbb{N}$, $f_k : \mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is differentiable. We define a *diagonal stochastic gradient descent (DSGD)* sequence:

$$\theta_{k+1} := \theta_k - \alpha_{k+1}\nabla_\theta f_{k+1}(\theta_k,\mathbf{z}_{k+1}) \qquad \textbf{(DSGD)}$$

where $\mathbf{z}_{k+1}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.

Assume $\alpha_k = \Theta(1/k)$ and that the $f_k$ converge (pointwise) to $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$. Let

$$g_k(\theta) := \mathbb{E}_\mathbf{z}[f_k(\theta,\mathbf{z})] \qquad g(\theta) := \mathbb{E}_\mathbf{z}[f(\theta,\mathbf{z})]$$

## Abstract Convergence

Suppose the $g_k$ and $g$ are well-defined and differentiable. Suppose there exist $\{\theta_i \mid i\in\mathbb{N}\}\subseteq\mathbf{\Theta}\subseteq\mathbb{R}^m$, $L>0$ and $\epsilon>0$ s.t. for all $k\in\mathbb{N}$ and $\theta\in\mathbf{\Theta}$,

1. $\nabla_\theta g_k(\theta) = \mathbb{E}_\mathbf{z}[\nabla_\theta f_k(\theta,\mathbf{z})]$ (unbiased)
2. $|g_{k+1}(\theta) - g_k(\theta)| < k^{-1-\epsilon}\cdot L$ (uniform convergence)
3. $\|\nabla g_k(\theta) - \nabla g(\theta)\|^2 < k^{-\epsilon}\cdot L$ (gradient uniform convergence)
4. $\mathbb{E}_\mathbf{z}[\|\nabla_\theta f_k(\theta,\mathbf{z})\|^2] < L$ (“variance” bounded)
5. $\|\mathbf{H}\, g_k(\theta)\| < L$ (Hessian bounded)

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

*Instantiate $f_k$ with $[\![M]\!]_k$*

## Diagonalisation Gradient Descent for Programs

Let $M$ be a term of type $R$ with free variables $\theta$ and $\mathbf{z}$ of type $R$.

$$\theta_{k+1} := \theta_k - \alpha_{k+1}\nabla_\theta\,[\![M]\!]_{k+1}(\theta_k,\mathbf{z}_{k+1}) \qquad \textbf{(DSGD')}$$

where $\mathbf{z}_{k+1}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.

If $\mathbf{\Theta} := \{\theta_i \mid i\in\mathbb{N}\}$ is bounded then the *conditions for convergence* are *satisfied*.

*Use **(DSGD')** to find stationary points of the optimisation problem.*

## References

[LYY18] Wonyeol Lee, Hangyeol Yu, and Hongseok Yang: Reparameterization gradient for non-differentiable models. NeurIPS 2018.

[MOPW21] Carol Mak, C.-H. Luke Ong, Hugo Paquet, Dominik Wagner: Densities of Almost Surely Terminating Probabilistic Programs are Differentiable Almost Everywhere. ESOP 2021.

dominik.wagner@cs.ox.ac.uk