NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

Dominik Wagner, Basim Khajwal and Luke Ong

paper:

Introduction

Objective: solve stochastic optimisation problems expressed in programming languages

Example: variational inference for probabilistic programming

Main Contribution:

- novel variant of SGD (**Diagonalisation SGD**) for **non-differentiable** models, which follows the reparameterisation gradient estimator on a **smooth** approximation whilst **enhancing the accuracy** in each step
- **provable** convergence to stationary points of the **unsmoothed** objective

Problem Statement

Idealised Programming Language:

$$F ::= z_j \mid f(F,\dots,F) \mid \textbf{if } F<0 \textbf{ then } F \textbf{ else } F$$

primitive operation

discontinuous

$$\boxed{\text{argmin}_{\theta\in\Theta} \ \mathbb{E}_{s\sim\mathcal{D}}\left[\,[\![F]\!]\,(\phi_\theta(s))\right]}$$

F: term in language, $\mathcal{D}$: continuous probability distribution on $\mathbb{R}^n$,
$\Theta\subseteq\mathbb{R}^m$: parameter space, each $\phi_\theta:\mathbb{R}^n\to\mathbb{R}^n$ is a diffeomorphism*

* with suitable assumptions guaranteeing the objective is well-defined

in practice: *apply **stochastic gradient descent***

Gradient Estimation

- **Score Estimator:** **widely applicable** but **high variance** in practice
- **Reparametrisation Estimator:**

$$\nabla_\theta \,[\![F]\!]\,(\phi_\theta(s)) \qquad \text{where } s\sim\mathcal{D}$$

typically **lower variance** but may be **biased!** [LYY18]

(Unbiasedness) $\nabla_\theta\,\mathbb{E}_{s\sim\mathcal{D}}\left[[\![F]\!]\,(\phi_\theta(s))\right] \overset{!}{=} \mathbb{E}_{s\sim\mathcal{D}}\left[\nabla_\theta\,[\![F]\!]\,(\phi_\theta(s))\right]$

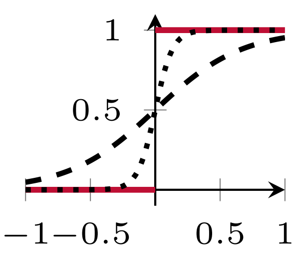
$$\boxed{\nabla_\theta\,\mathbb{E}_{s\sim\mathcal{N}(0,1)}\left[[\theta+s\geq 0]\right] \neq \mathbb{E}_{s\sim\mathcal{N}(0,1)}\Big[\underbrace{\nabla_\theta\,[\theta+s\geq 0]}_{=0\text{ a.e.}}\Big]}$$

Smoothing

interpret term F by smooth approximation $[\![F]\!]_\eta$ — accuracy coefficient

$$\boxed{\begin{aligned}[\![\textbf{if } F<0 \textbf{ then } G \textbf{ else } H]\!]_\eta(z) := {}&\sigma_\eta(-[\![F]\!]_\eta(z))\cdot[\![G]\!]_\eta(z)\\ &+\sigma_\eta([\![F]\!]_\eta(z))\cdot[\![H]\!]_\eta(z)\end{aligned}}$$

logistic sigmoid

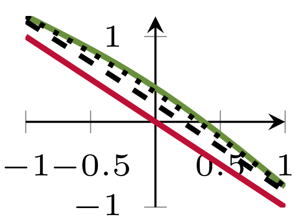


(Unbiasedness). ✓ (for each $\eta>0$)

$$\nabla_\theta\,\mathbb{E}_{s\sim\mathcal{D}}\left[[\![F]\!]\,(\phi_\theta(s))\right]_\eta = \mathbb{E}_{s\sim\mathcal{D}}\left[\nabla_\theta\,[\![F]\!]\,(\phi_\theta(s))\right]_\eta$$

idea: *apply SGD to smoothing*

quality of approximation?



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

(Uniform Convergence). Under mild syntactic assumptions,

$$\mathbb{E}_{s\sim\mathcal{D}}\left[[\![F]\!]_\eta(\phi_\theta(s))\right] \xrightarrow{\text{unif.}} \mathbb{E}_{s\sim\mathcal{D}}\left[[\![F]\!]\,(\phi_\theta(s))\right] \qquad \text{as } \eta\searrow 0$$

Counter-Example: For $F\equiv\textbf{if } 0<0 \textbf{ then } 0 \textbf{ else } 1$, $[\![F]\!]_\eta = 0.5 \not\to [\![F]\!] = 1$.

Diagonalisation SGD

Problem: *choice of accuracy coefficients η?*
Solution: *enhance accuracy coefficient in each step*
(rather than fixing η in advance)

$$\boxed{\theta_{k+1} := \theta_k - \gamma_k\cdot\nabla_\theta\,[\![F]\!]_{\eta_k}(\phi_{\theta_k}(s_k)) \qquad s_k\sim\mathcal{D}}$$

$(\gamma_k)_{k\in\mathbb{N}}$ step sizes, $(\eta_k)_{k\in\mathbb{N}}$ schedule of accuracy coefficients s.t. $\eta_k\searrow 0$.

Problem: variance grows as $\eta\searrow 0$
Solution:

- tame variance with suitable schedule of accuracy coefficients η_k
- bound growth of variance based on **syntactic** shape of expressions (nesting depth of conditionals into guards of if-statements)

Example:

nesting depth 1: $F_1\equiv -0.5\cdot z^2 + \textbf{if } z<0 \textbf{ then } 0 \textbf{ else } 1$

nesting depth 2: $F_2\equiv \textbf{if } (a\cdot(\textbf{if } b\cdot z_1+c<0 \textbf{ then } 0 \textbf{ else } 1) + d\cdot(\textbf{if } e\cdot z_2+f<0 \textbf{ then } 0 \textbf{ else } 1)+g)<0 \textbf{ then } 0 \textbf{ else } 1$

Theorem (Correctness of Diagonalisation SGD).

Suppose F has nesting depth ℓ of if-statements into guards and $\epsilon>0$. Then DSGD is correct for $\gamma_k\in\Theta(1/k)$ and $\eta_k\in\Theta(k^{-\frac{1}{\ell}+\epsilon})$: almost surely

$$\liminf_{i\to\infty} \left\|\nabla_{\theta_i}\,\mathbb{E}_{s\sim\mathcal{D}}\left[[\![F]\!]\,(\phi_{\theta_i}(s))\right]\right\| = 0$$

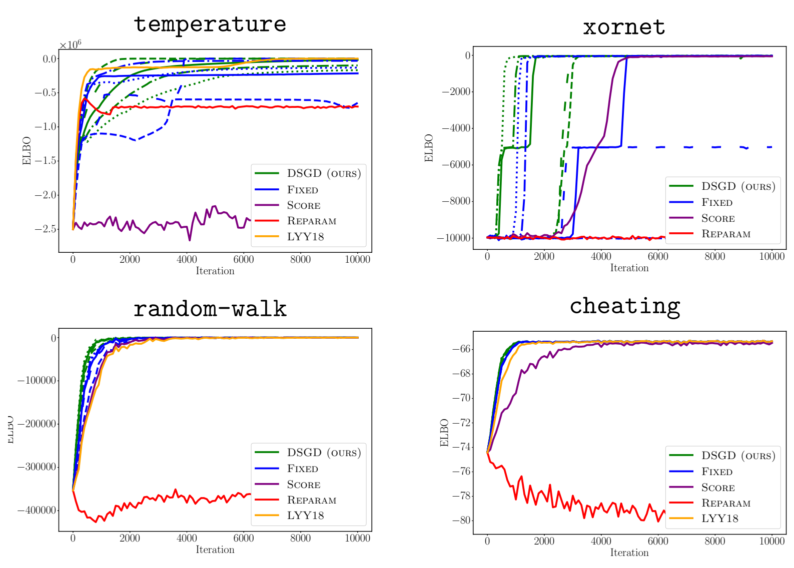
or $\theta_i\notin\Theta$ for some $i\in\mathbb{N}$.

Example: For nesting depth 1, $\eta_k\in\Theta(1/\sqrt{k})$ can be chosen.

*Only the **syntactic** structure of terms is essential for the choice of $(\eta_k)_{k\in\mathbb{N}}$!*

Empirical Evaluation

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**.



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

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
LYY18	not applicable		

Fixed uses smoothing with fixed accuracy coefficient $\eta=\eta_{4000}$ [KOW23].
LYY18 corrects bias of standard reparameterisation estimator (Reparam) by computing a boundary term [LYY18].

References

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

[KOW23] Basim Khajwal, C.-H. Luke Ong, Dominik Wagner: Fast and Correct Gradient-Based Optimisation for Probabilistic Programming via Smoothing. ESOP 2023.

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