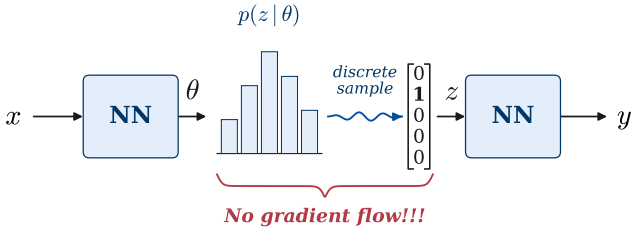
Generalized and Optimal Straight-Through Estimators

James David Hooper | Alexander Shekhovtsov

Czech Technical University in Prague

Introduction & Overview

- Many ML models contain **discrete random variables** sampled on the forward pass.
- Backpropagation is difficult: sampling has no well-defined derivative.

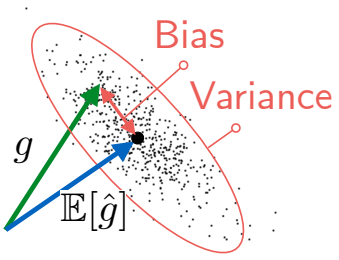


- **Many prior methods:** substitute a **surrogate Jacobian** during backprop (approximate chain rule).
- **This work:** a **unified framework** that axiomatically defines the class of *generalized straight-through* (GST) estimators and derives existing methods as special cases.
- **New estimators:** found via a **minimum-variance principle** – provably better than existing biased estimators.

The Problem: Bias vs. Variance

Let $x \in \mathcal{X}$ be a discrete sample on the forward pass. We want to estimate the gradient of the expected loss from a **single sample** x. The estimator is itself a **random variable**. Empirically, both **bias** and **variance** impact training.

- **Unbiased estimators** (REINFORCE):
  high variance – difficult to use in some applications.
- **Biased approximate chain rule estimators** (ST, Gumbel-Softmax, ZGR, ReinMax):
  lower variance – more useful in practice.
- **Approximate chain rule estimators** are derived from *different ad-hoc principles* and come with **no optimality guarantees**.



Motivation: Can we define the space of all sensible approximate chain rule estimators and find *optimal* estimators with respect to bias and variance?

Axiomatic Approach: Generalized ST (GST)

Axioms:

A1. Linearity

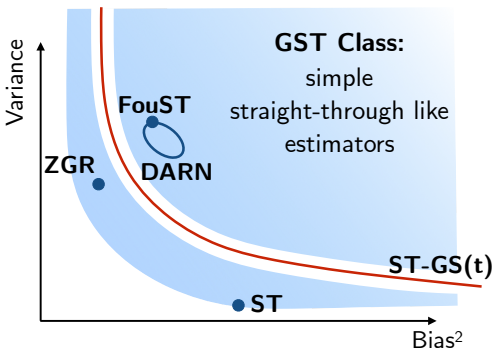
The estimator is linear in the loss function:

$$\hat{J}[\mathcal{L}_1+\mathcal{L}_2] = \hat{J}[\mathcal{L}_1] + \hat{J}[\mathcal{L}_2]$$

A2. Linear unbiasedness

Unbiased for any linear loss $\mathcal{L}(\phi) = a^\top \phi$:

$$\mathbb{E}[\hat{J}] = J$$



Minimum Variance Estimators (MVE)

Formalise the bias-variance trade-off as a **Lagrangian optimisation** within the GST class, then solve in closed form for the optimal estimator.

Hard MVE

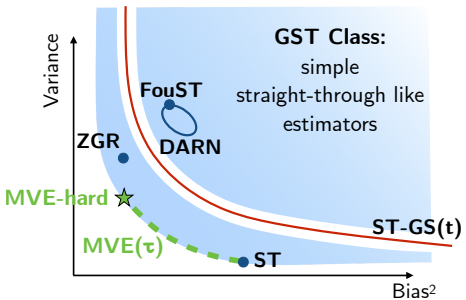
min variance proxy

s.t. zero linear bias , zero quadratic bias

Soft MVE(τ)

min variance proxy $+ \tau \cdot$ quadratic bias penalty

s.t. zero linear bias



Key Results & Conclusion

Unifying Existing Estimators

- Our framework cleanly demonstrates that ST, FouST, DARN, ZGR, and ReinMax are contained within the GST class.
- We formally re-derive ST and ZGR as optimal estimators in the GST class, and prove that **ReinMax $\equiv$ ZGR**.

New Optimal Estimators

- We derive a family of minimum variance GST estimators **(MVE)** in closed form for two practically important settings: the 1D case and the one-hot embedding.

MVE is competitive with or better than every existing estimator, across settings.