International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Generalized Gradient Estimation for Variational Autoencoders and Reinforcement Learning

Hyacinthe Hamon

Hamon FZCO, Research And Development

Abstract

This work presents a generalized gradient estimator that optimizes expectations involving known or black-box functions for discrete and continuous random variables. We synthesize and extend standard methods for constructing gradient estimators, offering a framework that incurs minimal computational overhead. Our proposed approach demonstrates effectiveness in variational autoencoders and introduces a straightforward extension to reinforcement learning, accommodating discrete and continuous action settings. Experiment-tal results reveal improved training performance and sample efficiency, highlighting the utility of our estimator in various domains. Future applications include training models with complex attention mechanisms, continuous latent-variable models with non-differentiable likelihoods, and integrating our method with existing variance-reduction techniques and optimization methods in reinforcement learning.

Keywords: Gradient Estimation, Variational Autoencoders (VAEs), Reinforcement Learning, Reparameterization Trick, Control Variates, Policy Gradient Methods

1. INTRODUCTION

Slope-based enhancement supports propels in AI and support learning. Backpropagation [16,19,12] figures definite slopes for differentiable targets, while the reparameterization trick [24,4,13] empowers the practical improvement of probabilistic models.

Notwithstanding, numerous targets need slopes for backpropagation, for example, black-box capabilities in support learning [18] or discontinuities from discrete sampling [7,2]. Ongoing techniques address this with angle assessors, including entertainer pundit methods [21] and persistent relaxations [7,2]. [22] presented a fair-minded, low-fluctuation assessor through persistent relaxations. We broaden this by learning a brain network-based control variate, yielding a lower-fluctuation, fair assessor material even without consistent relaxations, as in support learning or black-box improvement.

2. BACKGROUND

2.1. Inclination Estimators

Streamlining boundaries θ to expand an assumption shows up in support learning (expected reward $E\tau\sim\pi[R]$) and dormant variable models (boosting $p(x|\theta) = E_{p(z|\theta)}[p(x|z)]$). We enhance

$L(\theta) = E_{p(b|\theta)}[\, f\,(b)]$.

(1)

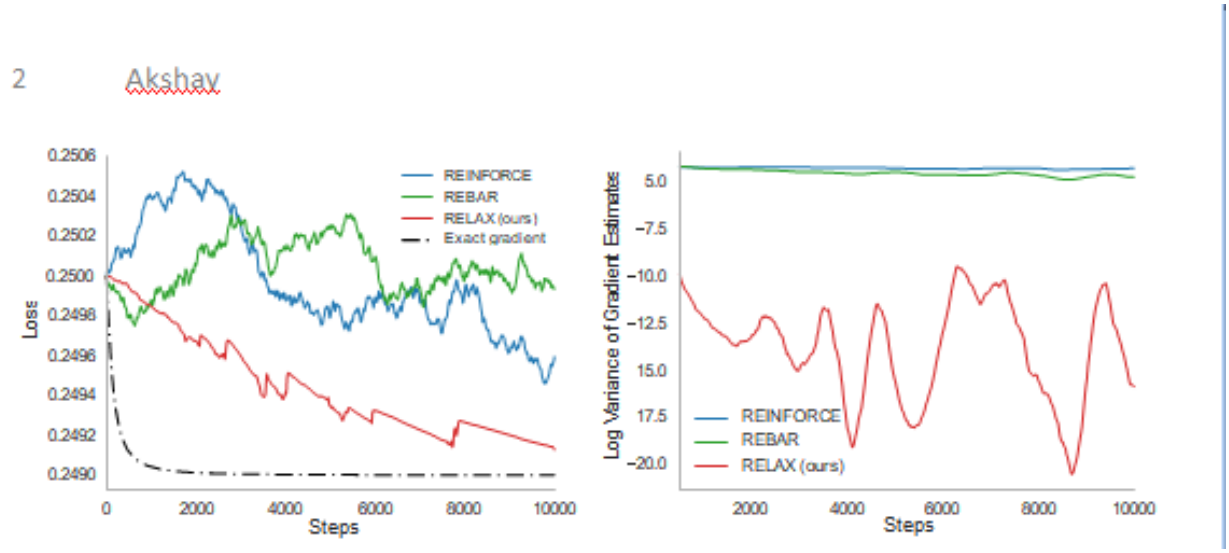


Figure 1: Left: Training Curves Comparing Different Gradient Estimators on a Toy Problem: $L(\Theta) = E_{p(B|\Theta)}[(B - 0.499)^2]$ Right: Log-Variance Of Each Estimator's Gradient.

For high-layered θ, unprejudiced stochastic inclinations gˆ are required for convergence [14].
Key inclination assessment strategies:
Score-capability assessor (REINFORCE) [24] is fair-minded yet high-fluctuation:

$$\cap g_{REINFORCE}[f] = f(b)\nabla \log p(b|\theta), \qquad b \sim p(b|\theta) \quad (2)$$

Reparameterization trick [24,4,13] lessens difference by communicating *b* as an element of an irregular variable:

$$\cap \text{greparam}[f] = \frac{\partial f}{\partial b}\frac{\partial b}{\partial \theta} \qquad \epsilon \sim p(\epsilon) \qquad (3)$$

Control variates [14] lessen difference by deducting a known-capability c(b):

$$\cap_{new}(b) = g\hat{}(b) - c(b) + E_{p(b)}[c(b)] \qquad (4)$$

This brings down change if c(b) is related to gˆ(b), with additional enhancement through a learned scalar [22].

3. CONSTRUCTING AND UPGRADING A DIFFERENTIABLE SURROGATE

We present an inclination assessor for the assumption ∂ Ep(b|θ)[f (b)] utilizing a mix of score capability assessors, the reparameterization stunt, and control variates. For persistent b where f isn't differentiable, we develop a proxy cϕ
utilizing a brain organization and separate through it. The assessor, Careless, is:

$$\cap g_{LAX} = g\hat{}_{REINFORCE}[f] - g\hat{}_{REINFORCE}[c_\phi] + g\hat{}_{reparam}[c_\phi]$$

$$= [f(b) - c\phi(b)]\frac{\partial}{\partial\theta} log p(b|\theta) + \frac{\partial}{\partial\theta} c\phi(b). \qquad (5)$$

This assessor is fair and can accomplish lower fluctuation than the reparame- terization assessor. Improvement of cϕ is done through inclination plunge, with the angle of the difference registered utilizing:

$$\text{Variance}(g\hat{}) = E\frac{\partial}{\partial\phi} g\hat{}2\,. \qquad (6)$$

Algorithm ?? Frames the streamlining of both θ and ϕ. For discrete factors, we present a casual variable z and apply the reparameterization stunt, prompting the assessor:

$$\cap g_{DLAX} = f(b)\frac{\partial}{\partial\theta} log p(b|\theta) - c\phi(z)\frac{\partial}{\partial\theta} log p(z|\theta) + \frac{\partial}{\partial\theta} c\phi(z), \quad b = H(z), z \sim p(z|\theta). \quad (7)$$

Further refinements lead to the Loosen up assessor, which involves a restrictive reparameterization for decreased difference:

$$\cap g_{RELAX} = f(b) - c\phi(\tilde{z})]\frac{\partial}{\partial\theta} log p(b|\theta) + \frac{\partial}{\partial\theta} c\phi(z) - \frac{\partial}{\partial\theta} c\phi(\tilde{z}), \quad b = H(z), z \sim p(z|\theta), \; \tilde{z} \sim p(z|\theta), \tilde{z} \sim p(z|b,\theta). \quad (8)$$

In support learning, we upgrade strategies utilizing a comparable slope assessor that consolidates a benefit capability:

$$A_t = \sum_{t'=t}^{T} r_t' - c_\phi(a_t, s_t), \quad (9)$$

$$= \sum_{t'=1}^{T} \partial log \pi_{LAX}(a|s, \vartheta \; " \; \sum_{t'=t}^{T} r_t' - c_\phi(a_t, s_t), + \frac{\partial}{\partial\theta} c_\phi(a_t, s_t). \quad (10)$$

4. SCOPE AND LIMITATIONS

Our work is firmly connected with the REBAR method [22], which is an excerptanational instance of the RELAX assessor, with the substitute set to cϕ(z) = η • *f* (softmaxλ(z)).

REBAR has restricted scope for improvement because of its dependence on the scaling factor η and temperature λ and must be applied when f is known and differentiable. It likewise relies upon unclear behavior, assessing discrete misfortune capabilities at nonstop data sources.

Conversely, LAX and RELAX can enhance black-box capabilities, for example, in Support picking up, requiring the capacity to question and separate p(b|θ).

Direct reliance on parameters When f relies upon θ, like in probabilistic models or with regularizers, we can broaden the slope assessors by adding ∂ f (b, θ), prompting an impartial assessor:

$$\frac{\partial}{\partial\theta} E_{p(b|\theta)}[f(b,\theta)] = E_{p(b|\theta)} + \frac{\partial}{\partial\theta} f(b,\theta) \; \frac{\partial}{\partial\theta} \log p(b|\theta) \quad (11)$$

5. RELATED WORK

A few late works have zeroed in on slope assessment. [8] lessen the difference of reparameterization slopes involving a basic model as a control variate. NVIL [9] and VIMCO [10] diminish change in discrete idle variable models, while [17] utilize limited contrasts to gauge angles in equal. Our strategy, notwithstanding, is a solitary example assessor.

[20] and [23] additionally use examining appropriations to construct angle assessors. [11] presents a non-parametric control variate to decrease the difference in the Monte Carlo mix. Late work on activity subordinate baselines in support learning, for example,

[1] and [6] share similitudes with LAX in ceaseless control assignments, with [25] utilizing per-aspect freedom for activity subordinate fair-minded baselines.

6. APPLICATIONS

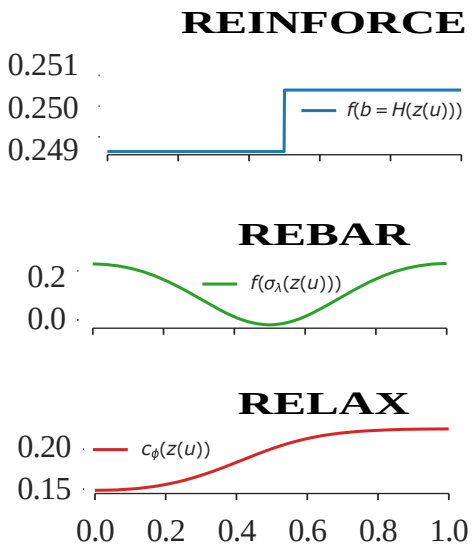
We show the viability of our assessor on different improvement issues, beginning with a toy model, trailed by streamlining twofold VAEs and support learning.

6.1. TOY EXPERIMENT

We limit $E_{p(b|\theta)}[(b - t)2]$ where $p(b|\theta) = \text{Bernoulli}(b|\theta)$, and $t = 0.499$, as recommended by [22]. Figures 1a and 1b show the general presentation and inclination log fluctuation of Build, REBAR, and Unwind. Figure 2 shows that the learned proxy $c\phi$ intently approximates f for all z, guaranteeing a little difference for Support and inclinations that improve the assumption. Conversely, REBAR's substitute approximates f close to 0 and 1, prompting less compelling enhancement. Consequently, Unwind accomplishes the best exhibition.

6.2. DISCRETE VARIATIONAL AUTOENCODER

We assess RELAX on preparing a variational autoencoder [4,13] with Bernoulli idle factors on MNIST and Omniglot [5]. We utilize the control variate $c_\phi(z) = f(\sigma\lambda(z)) + r\hat{}_\rho(z)$, where $r\hat{}_\rho$ is a brain organization and $f(\sigma\lambda(z))$ is the ELBO assessed at ceaseless contributions, as in REBAR.



u..

Figure 2: Ideal unwinding for a toy misfortune capability utilizing different inclination assessors: REBAR utilizes a quadratic substantial unwinding, while Unwind considers a freestyle unwinding. The learned control variate was further developed in all tests to prepare execution, contrasted with REBAR, especially in direct models, and accomplish quicker combinations. The decline in execution for nonlinear models is ascribed to overfitting. We will pass on additional investigation of this to future work.

Preparing bends in Figure 3 and Table?? This shows that RELAX prompts quicker union in both direct models than REBAR.

Dataset	Model	Concrete	NVIL	MuProp	REBAR	RELAX
MNIST	Nonlinear	−102.2	−101.5	−101.1	-81.01	**-78.13**
	Single Layer Linear	-111.3	−112.5	−111.7	-111.6	**-111.20**
	Double Layer Linear	-99.62	−99.6	−99.07	-98.22	**-98.00**
Omniglot	Nonlinear	−110.4	−109.58	−108.72	-56.76	**-56.12**
	Single Layer Linear	-117.23	−117.44	−117.09	-116.63	**-116.57**
	Double Layer Linear	-109.95	−109.98	−109.55	-108.71	**-108.54**

Table 1: Maximum training ELBO for discrete variational autoencoders.

6.3. REINFORCEMENT LEARNING

We apply RELAX and LAX assessors to support learning undertakings with discrete and ceaseless activities, contrasting them with A2C [21]. In discrete undertakings (Truck Shaft, Lunar Lander), we discard reward bootstrapping,

6 Akshay

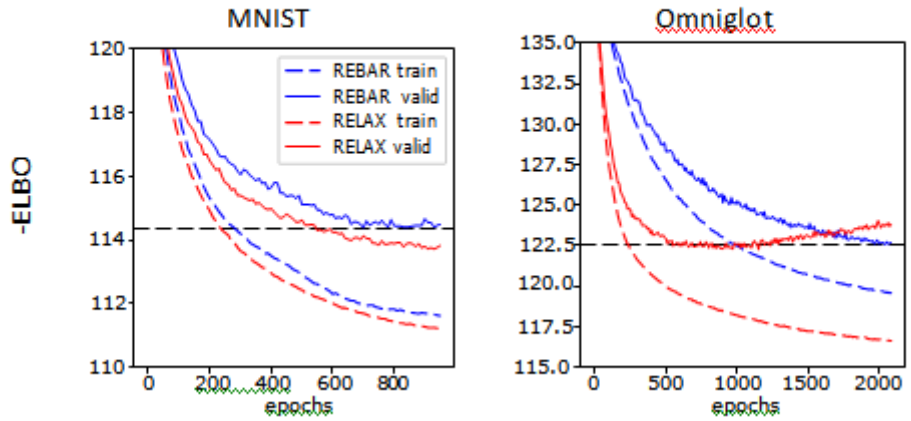


Figure 3: Preparing bends for one-layer direct VAE, with a ran line showing REBAR's most reduced approval blunder.

While performing consistent errands (Rearranged Pendulum), we utilize a worthwhile capability for bootstrapping. The control variate is $c_\phi(a, s) = V(s) + c\hat{}(a, s)$.

Our assessor lessens the inclination to change, permitting more considerable learning rates and quicker assembly. In discrete undertakings, we accomplish more than a two-times speedup over A2C. Results are displayed in Figure 4 and Table**.??**

Table 2: Average Episodes Needed To Complete Tasks. Task Completion Criteria Are Detailed In Appendix 12.

Approach	**Cart-Pole Lunar Lander**	**Inverted Pendulum**
A2C	1152 ± 90	162, 374 ± 17, 241	6, 243 ± 164
LAX/RELAX	**472 ± 114**	**68, 712 ± 20, 668**	**2, 067 ± 412**

7. CONCLUSIONS AND FUTURE WORK

We introduced a summed-up slope assessor with negligible computational overhead for assumptions of known or discovered elements of discrete or persistent irregular factors. We extended it to support learning for both discrete and continuous action spaces.

Future work includes applying our assessors to models with intricate attention or memory indexing [27], non-differentiable latent variable models such as 3D rendering engines, and extending reparameterization gradient estimators [15,?]. In reinforcement learning, integrating our approach with variance reduction techniques like generalized advantage estimation [3,?], optimization methods like KFAC [26], and off-policy algorithms like Q-prop [1] presents a promising direction.

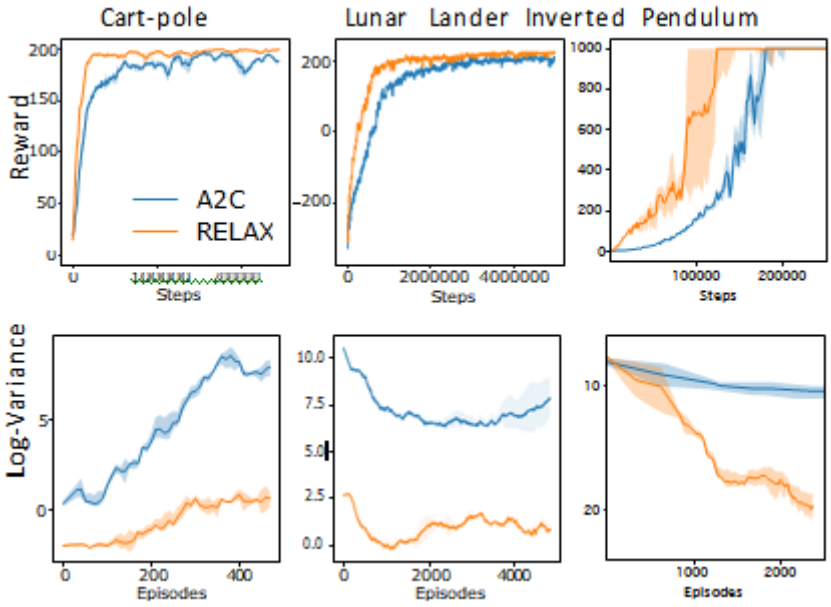


Figure 4: Top row: Reward curves. Bottom row: Log-fluctuation of strategy angles per episode, showing mean prizes (focus line) and changeability (bars). The middle value of over-addressed boundaries was found in each tenth episode.

References

1. Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R.E., Levine, S.: Q-prop: Sample- efficient policy gradient with an off-policy critic. arXiv preprint arXiv:1611.02247 (2016)
2. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
3. Kimura, H., Kobayashi, S., et al.: An analysis of actor-critic algorithms using eligibility traces: reinforcement learning with imperfect value functions. Journal of Japanese Society for Artificial Intelligence 15(2), 267–275 (2000)
4. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. International Conference on Learning Representations (2014)
5. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science 350(6266), 1332–1338 (2015)
6. Liu, H., Feng, Y., Mao, Y., Zhou, D., Peng, J., Liu, Q.: Sample-efficient policy optimization with stein control variate. arXiv preprint arXiv:1710.11198 (2017)

7. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712 (2016)
8. Miller, A.C., Foti, N.J., D'Amour, A., Adams, R.P.: Reducing reparameterization gradient variance. arXiv preprint arXiv:1705.07880 (2017)
9. Mnih, A., Gregor, K.: Neural variational inference and learning in belief networks. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1791–1799 (2014)
10. Mnih, A., Rezende, D.: Variational inference for Monte Carlo objectives. In: International Conference on Machine Learning. pp. 2188–2196 (2016)
11. Oates, C.J., Girolami, M., Chopin, N.: Control functionals for Monte Carlo integration. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(3), 695–718 (2017)
12. Rall, L.B.: Automatic differentiation: Techniques and applications (1981)
13. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on Machine Learning. pp. 1278–1286 (2014)
14. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics pp. 400–407 (1951)
15. Ruiz, F.R., AUEB, M.T.R., Blei, D.: The generalized reparameterization gradient. In: Advances in Neural Information Processing Systems. pp. 460–468 (2016)
16. Rumelhart, D.E., Hinton, G.E.: Learning representations by back-propagating errors. Nature 323, 9 (1986)
17. Salimans, T., Ho, J., Chen, X., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864 (2017)
18. Schulman, J., Heess, N., Weber, T., Abbeel, P.: Gradient estimation using stochastic computation graphs. In: Advances in Neural Information Processing Systems. pp. 3528–3536 (2015)
19. Speelpenning, B.: Compiling Fast Partial Derivatives of Functions Given by Algo- rithms. Ph.D. thesis, University of Illinois at Urbana-Champaign (1980)
20. Staines, J., Barber, D.: Variational optimization. arXiv preprint arXiv:1212.4507 (2012)
21. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems. pp. 1057–1063 (2000)
22. Tucker, G., Mnih, A., Maddison, C.J., Sohl-Dickstein, J.: Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. arXiv preprint arXiv:1703.07370 (2017)
23. Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., Schmidhuber, J.: Natural evolution strategies. Journal of Machine Learning Research 15(1), 949–980 (2014)
24. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning 8(3-4), 229–256 (1992)
25. Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A.M., Kakade, S., Mor- datch, I., Abbeel, P.: Variance reduction for policy gradient with action-dependent factorized baselines. International Conference on Learning Representations (2018)
26. Wu, Y., Mansimov, E., Liao, S., Grosse, R., Ba, J.: Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. In: Advances in neural information processing systems (2017)
27. Zaremba, W., Sutskever, I.: Reinforcement learning neural turning machines-revised. arXiv preprint

arXiv:1505.00521 (2015)