Stochastic Variance-Reduced Policy Gradient

Matteo Papini
Damiano Binaghi   Giuseppe Canonaco
Matteo Pirotta   Marcello Restelli

35th International Conference on Machine Learning, Stockholm, Sweden

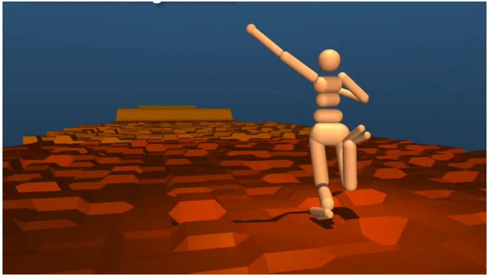
Outline

Stochastic Variance-Reduced (Policy) Gradient

- **SVRG** for Reinforcement Learning
  - Motivation
  - Challenges
- **SVRPG**
  - Convergence Properties
  - Heuristics
  - Experiments

Policy Gradient

An effective **Reinforcement Learning (RL)** solution to **continuous** control problems:

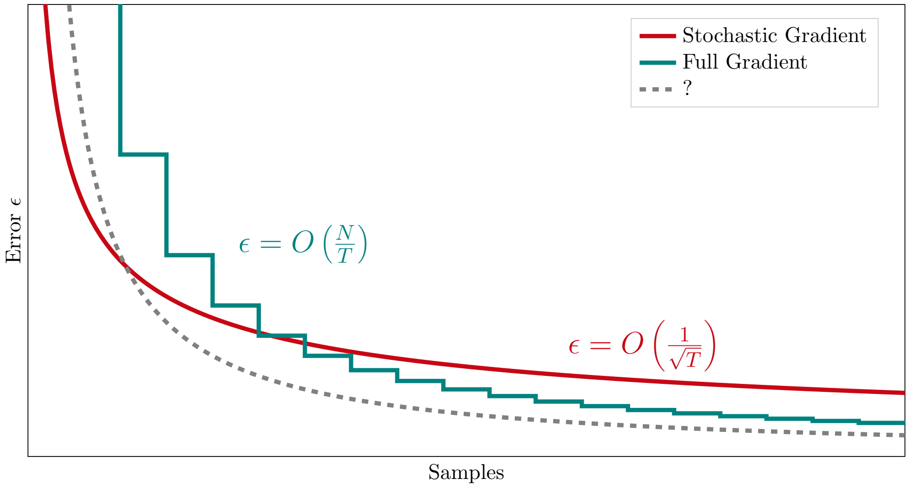


Robotics (Heess et al., 2017)

Video games (OpenAI, 2018)

Mostly based on **Stochastic Gradient Ascent** (Robbins and Monro, 1951)

Full vs Stochastic Gradient
Non-convex optimization



Can we do something better?

A solution from **finite-sum optimization**:

$$\underbrace{\blacktriangledown J(\boldsymbol{\theta})}_{\text{SVRG estimator}} = \underbrace{\nabla J(\widetilde{\boldsymbol{\theta}})}_{\text{FG in snapshot parameter}} + \underbrace{\nabla J(\boldsymbol{\theta})\big|_{\tau_i}}_{\text{SG in current parameter}} - \underbrace{\nabla J(\widetilde{\boldsymbol{\theta}})\big|_{\tau_i}}_{\text{Correction term}}$$

- Unbiased
- Linear convergence
- More data-efficient than FG
- Easily applicable to **Supervised Learning (SL)**

Not trivial! There are three **challenges**:

1. **Non-concavity** of $J(\boldsymbol{\theta})$ (Allen-Zhu and Hazan, 2016; Reddi et al., 2016)
2. **Infinite dataset**: we would need *infinite samples* to compute FG (Harikandeh et al., 2015; Bietti and Mairal, 2017)
3. **Non-stationarity**: $\tau \sim p_{\boldsymbol{\theta}}$ (new!)

RL so far: *policy evaluation* (Du et al., 2017) and *off-policy control* (Xu et al., 2017)

Our work: **on-policy control**

$$\underbrace{\blacktriangledown J(\boldsymbol{\theta})}_{\text{SVRPG estimator}} = \underbrace{\widehat{\nabla}_N J(\widetilde{\boldsymbol{\theta}})}_{\substack{\text{Large N}\\ \text{to approximate FG}}} + \underbrace{\widehat{\nabla}_B J(\boldsymbol{\theta})}_{B<<N} - \underbrace{\omega(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})\widehat{\nabla}_B J(\widetilde{\boldsymbol{\theta}})}_{\substack{\text{Importance weighting}\\ \text{for non-stationarity}}}$$

- Unbiased
- More data-efficient than FG

Convergence Properties

Convergence to **local** optimum:

$$\mathbb{E}\left[\left\|\nabla J(\boldsymbol{\theta})\right\|^2\right] \leq \frac{J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0)}{\psi T} + \underbrace{\frac{\zeta}{N}}_{\text{Infinite dataset}} + \underbrace{\frac{\xi}{B}}_{\text{Nonstationarity}}$$

- Linear convergence + error (similar to Harikandeh et al., 2015)

Heuristics

Meta-parameter selection

- **Adaptive step size**: two ADAM annealing schedules

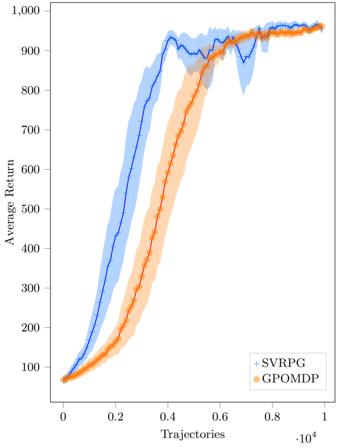
$$\underbrace{\alpha_{FG}}_{\text{used at the snapshot}} \qquad \underbrace{\alpha_{SG}}_{\text{used inside epoch}}$$

- **Adaptive epoch size**: take new snapshot when the effective step size becomes too small

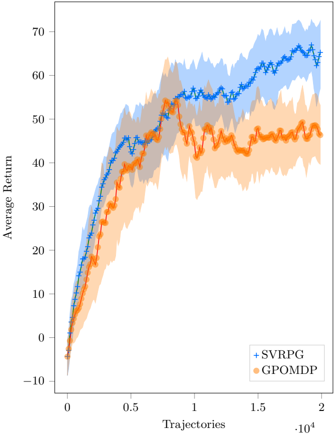
$$\frac{\alpha_{SG}}{B} < \frac{\alpha_{FG}}{N} \implies \textbf{snapshot}$$

Results

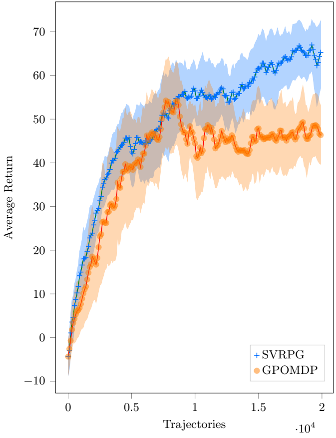
Cart-Pole



Swimmer



Half-Cheetah



Conclusions

- Efficient policy optimization is challenging
- **SVRPG**: on-policy control based on SVRG
- Meta-parameters still crucial to tame different sources of variance
- Future work: adaptive batch size, natural gradient, actor-critic

Thank you for your attention

- Poster: today 06:15 – 09:00 PM @ **Hall B #65**
- Contact: `matteo.papini@polimi.it`
- Online resources: `t3p.github.io`

References

Allen-Zhu, Z. and Hazan, E. (2016). Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707.

Bietti, A. and Mairal, J. (2017). Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, pages 1622–1632.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1049–1058. PMLR.

Harikandeh, R., Ahmed, M. O., Virani, A., Schmidt, M., Konečǹy, J., and Sallinen, S. (2015). Stopwasting my gradients: Practical svrg. In *Advances in Neural Information Processing Systems*, pages 2251–2259.

Heess, N., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, A., Riedmiller, M., et al. (2017). Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.

OpenAI (2018). Openai five.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Xu, T., Liu, Q., and Peng, J. (2017). Stochastic variance reduction for policy gradient estimation. *CoRR*, abs/1710.06034.

Adaptive Step Size

Adaptive Epoch Size

Normalized Importance Weights

Actor-Critic SVRPG

The Full Comparison