UNIVERSITY OF CAMBRIDGE

Monte Carlo gradient estimation in ML

James Urquhart Allingham

The paper...

Monte Carlo Gradient Estimation in Machine Learning

Shakir Mohamed[*1] SHAKIR@GOOGLE.COM
Mihaela Rosca[*1 2] MIHAELACR@GOOGLE.COM
Michael Figurnov[*1] MFIGURNOV@GOOGLE.COM
Andriy Mnih[*1] AMNIH@GOOGLE.COM
[*]*Equal contributions;*
1 *DeepMind, London*
2 *University College London*

Talk outline

- Problem setup
- Building some intuition
- 3 classes of gradient estimators and their properties:
  - pathwise
  - score function
  - measure valued
- Variance reduction techniques

The problem

Consider a probabilistic objective function $\mathcal{F}$:

$$\mathcal{F}(\boldsymbol{\theta}) := \int p(\mathbf{x}; \boldsymbol{\theta}) f(\mathbf{x}; \phi) \mathrm{d}\mathbf{x} = \mathbb{E}_{p(\mathbf{x}; \boldsymbol{\theta})} \left[f(\mathbf{x}; \phi) \right]$$

with a *cost f* and and *measure p*.
If we want to optimise this with respect to the distributional parameters $\boldsymbol{\theta}$, we must evaluate the gradient $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} := \nabla_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x}; \boldsymbol{\theta})} \left[f(\mathbf{x}; \phi) \right].$$

The challenge

Evaluating $\boldsymbol{\eta}$ is difficult

$$\boldsymbol{\eta} := \nabla_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x};\phi)\right]$$

- can't evaluate the expectation $\mathcal{F}(\boldsymbol{\theta})$ in closed form
- $\mathbf{x}$ is high dimensional – quadrature is ineffective
- $\boldsymbol{\theta}$ is high dimensional
- f is non-differentiable/black-box/expensive to evaluate

Monte Carlo estimators

We can solve the 1st problem by approximating $\mathcal{F}(\boldsymbol{\theta})$ as:

$$\bar{\mathcal{F}}_N = \frac{1}{N}\sum_{n=1}^{N} f\left(\hat{\mathbf{x}}^{(n)}\right), \quad \hat{\mathbf{x}}^{(n)} \sim p(\mathbf{x};\boldsymbol{\theta}).$$

This is a very general solution! 4 desired properties:

- Consistency

$$\lim_{N\to\infty} \bar{\mathcal{F}}_N = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x};\phi)\right]$$

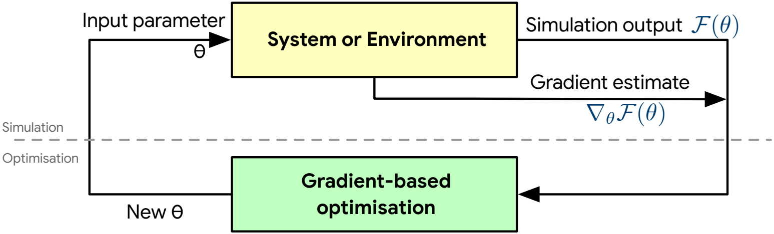
- Unbiasedness

$$\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\bar{\mathcal{F}}_N\right] = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right]$$

- Low variance $\mathbb{V}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\bar{\mathcal{F}}_N\right]$
- Efficiency

UNIVERSITY OF CAMBRIDGE

Stochastic Optimisation



Variational Inference

Here our objective has the same form as $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x};\phi)\right]$:

Variational Free Energy

$$\boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}\left[\log p(\mathbf{x}|\mathbf{z};\phi) - \log \frac{q(\mathbf{z}|\mathbf{x};\theta)}{p(\mathbf{z})}\right]$$

- model/likelihood $p(\mathbf{x}|\mathbf{z};\phi)$
- variational family $q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$
- prior $p(\mathbf{z})$

Model-free Reinforcement Learning

Once again we have an objective of the form $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})} \left[f(\mathbf{x}; \phi) \right]$:

Policy gradient

$$\boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{\tau};\boldsymbol{\theta})} \left[\sum_{t=0}^{T} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

- trajectories $\boldsymbol{\tau} = (\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \ldots, \mathbf{s}_T, \mathbf{a}_T)$
- $p(\boldsymbol{\tau}; \boldsymbol{\theta}) = \left[\prod_{t=0}^{T-1} p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) p(\mathbf{a}_t | \mathbf{s}_t; \boldsymbol{\theta}) \right] p(\mathbf{a}_T | \mathbf{s}_T; \boldsymbol{\theta})$

Other applications

Many other interesting and important problems boil down to optimisation of an objective like $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x};\phi)\right]$:

- sensitivity analysis (e.g. Black-Scholes option pricing model)
- discrete event systems and queuing theory
- experimental design

Looking specifically at ML applications:

- stochastic differential equations
- learning deep generative models
- bandits
- many more

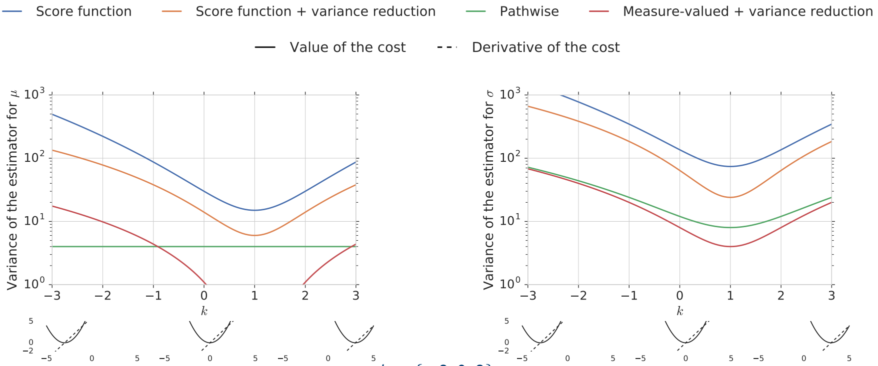
Building Intuition I

$$\eta = \nabla_\theta \int \mathcal{N}(x|\mu, \sigma^2) f(x; k) dx; \quad \theta \in \{\mu, \sigma\};$$

$$f \in \{(x - k)^2, \exp(-kx^2), \cos(kx)\}$$

Building Intuition II

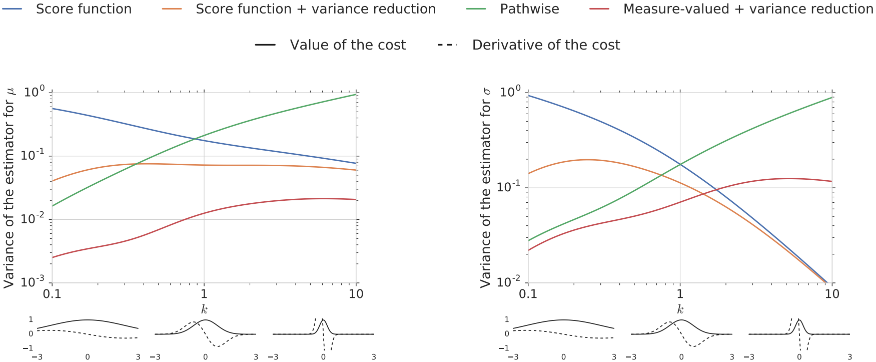
$$\nabla_\theta \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[(x-k)^2\right] \text{ for } \mu = \sigma = 1$$



$k \in \{-3, 0, 3\}$

Building Intuition III

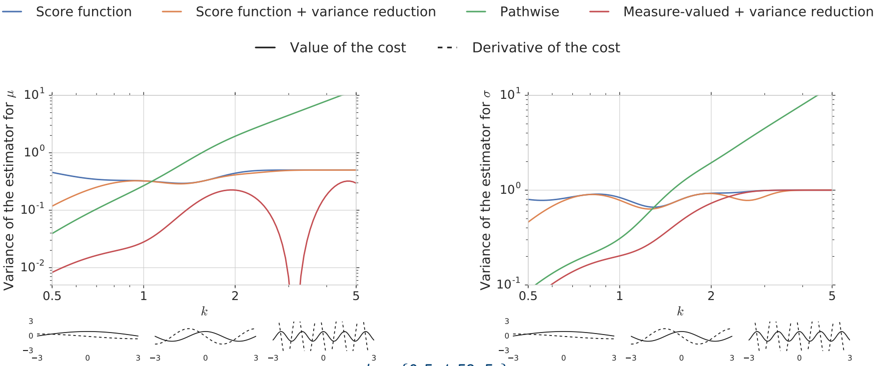
$$\nabla_\theta \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[\exp(-kx^2)\right] \text{ for } \mu = \sigma = 1$$



$k \in \{0.1, 1, 10\}$

Building Intuition IV

$$\nabla_\theta \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}[\cos kx] \text{ for } \mu = \sigma = 1$$



$k \in \{0.5, 1.58, 5.\}$

Score Function Estimator

Score Function

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\theta})}$$

Several useful properties:

- key quantity in MLE
- zero expectation:

$$\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta})\right] = \int p(\mathbf{x}; \boldsymbol{\theta}) \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\theta})} d\mathbf{x}$$
$$= \nabla_{\boldsymbol{\theta}} \int p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0}$$

- Variance is Fisher information

UNIVERSITY OF CAMBRIDGE

Score Function Estimator – Derivation

$$\eta = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right] = \nabla_{\boldsymbol{\theta}} \int p(\mathbf{x};\boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) \nabla_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}) d\mathbf{x}$$

$$= \int p(\mathbf{x};\boldsymbol{\theta}) f(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) d\mathbf{x}$$

$$= \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})\right]$$

$$\bar{\eta}_N = \frac{1}{N} \sum_{n=1}^{N} f(\hat{\mathbf{x}}^{(n)}) \nabla_{\boldsymbol{\theta}} \log p(\hat{\mathbf{x}}^{(n)};\boldsymbol{\theta}); \quad \hat{\mathbf{x}}^{(n)} \sim p(\mathbf{x};\boldsymbol{\theta})$$

Baseline corrected:

$$\eta = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[(f(\mathbf{x}) - \beta) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})\right]$$

Score Function Estimator – Unbiasedness

The score function estimator is unbiased if interchanging the integral and derivative is valid. Sufficient conditions are:

- $p(\mathbf{x};\boldsymbol{\theta})$ is continuously differentiable in its parameters $\boldsymbol{\theta}$.
- $f(\mathbf{x})p(\mathbf{x};\boldsymbol{\theta})$ is both integrable and differentiable for all parameters $\boldsymbol{\theta}$.
- There exists an integrable function $g(\mathbf{x})$ such that $\sup_{\boldsymbol{\theta}} \|f(\mathbf{x})\nabla_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})\|_1 \leq g(\mathbf{x})\ \forall \mathbf{x}$.

Score Function Estimator – Absolute Continuity

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{p(\mathbf{x};\theta)}\left[f(\mathbf{x})\right] &= \int \nabla_\theta p(\mathbf{x};\theta) f(\mathbf{x}) d\mathbf{x} \\
&= \lim_{h\to 0} \int \frac{p(\mathbf{x};\theta+h) - p(\mathbf{x};\theta)}{h} f(\mathbf{x}) d\mathbf{x} \\
&= \lim_{h\to 0} \frac{1}{h} \int p(\mathbf{x};\theta) \frac{p(\mathbf{x};\theta+h) - p(\mathbf{x};\theta)}{p(\mathbf{x};\theta)} f(\mathbf{x}) d\mathbf{x} \\
&= \lim_{h\to 0} \frac{1}{h} \int p(\mathbf{x};\theta) \left(\frac{p(\mathbf{x};\theta+h)}{p(\mathbf{x};\theta)} - 1 \right) f(\mathbf{x}) d\mathbf{x} \\
&= \lim_{h\to 0} \tfrac{1}{h} \left(\mathbb{E}_{p(\mathbf{x};\theta)}\left[\omega(\theta,h) f(\mathbf{x})\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[f(\mathbf{x})\right] \right)
\end{aligned}
$$

$\omega(\theta, h)$ implies an implicit assumption about absolute continuity. Violated for $\mathcal{U}[0, \theta]$.

Score Function Estimator – Variance

$$\mathbb{V}_{p(\mathbf{x};\theta)}[\bar{\eta}_N] = \mathbb{E}_{p(\mathbf{x};\theta)}\left[\left(f(\mathbf{x})\nabla_\theta \log p(\mathbf{x};\theta)\right)^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$
$$= \lim_{h\to 0} \frac{1}{h}\mathbb{E}_{p(\mathbf{x};\theta)}\left[(\omega(\theta,h)-1)^2 f(\mathbf{x})^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$

3 sources of variance:

1. Importance ratio ω:

$$\mathbb{E}_{p(\mathbf{x};\theta)}\left[(\omega(\theta,h)-1)^2 f(\mathbf{x})^2\right]$$

2. Dimensionality of $\mathbf{x}$
3. Cost function $f(\mathbf{x})$

Score Function Estimator – Variance

$$\mathbb{V}_{p(\mathbf{x};\theta)}[\bar{\eta}_N] = \mathbb{E}_{p(\mathbf{x};\theta)}\left[\left(f(\mathbf{x})\nabla_\theta \log p(\mathbf{x};\theta)\right)^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$
$$= \lim_{h\to 0} \frac{1}{h}\mathbb{E}_{p(\mathbf{x};\theta)}\left[(\omega(\theta,h)-1)^2 f(\mathbf{x})^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$

3 sources of variance:

1. Importance ratio ω
2. Dimensionality of $\mathbf{x}$:

$$\prod_{d=1}^{D} \mathbb{E}_{p(x_d;\theta)}\left[\frac{p(x_d;\theta+h)}{p(x_d;\theta)}\right] = 1, \quad \forall D$$

$$\lim_{d\to\infty} \omega(\theta,h) = \lim_{d\to\infty} \prod_d \frac{p(x_d;\theta+h)}{p(x_d;\theta)} = 0$$

3. Cost function $f(\mathbf{x})$

UNIVERSITY OF CAMBRIDGE

Score Function Estimator – Variance

$$\mathbb{V}_{p(\mathbf{x};\theta)}[\bar{\eta}_N] = \mathbb{E}_{p(\mathbf{x};\theta)}\left[\left(f(\mathbf{x})\nabla_\theta \log p(\mathbf{x};\theta)\right)^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$

$$= \lim_{h\to 0} \frac{1}{h}\mathbb{E}_{p(\mathbf{x};\theta)}\left[(\omega(\theta,h)-1)^2 f(\mathbf{x})^2\right] - \mathbb{E}_{p(\mathbf{x};\theta)}\left[\bar{\eta}_N\right]^2$$

3 sources of variance:

1. Importance ratio ω
2. Dimensionality of $\mathbf{x}$
3. Cost function $f(\mathbf{x})$:
   e.g. $f(\mathbf{x}) = \sum_k f(x_d)$, $\mathbb{V}[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) f(\mathbf{x})]$ will be of $O(D^2)$

UNIVERSITY OF CAMBRIDGE

Score Function Estimator – Computation

$$\eta = \nabla_\theta \mathbb{E}_{p(\mathbf{x};\theta)}[f(\mathbf{x})] = \mathrm{Cov}[f(\mathbf{x}), \nabla_\theta \log p(\mathbf{x};\theta)],$$
$$\mathrm{Cov}[f(\mathbf{x}), \nabla_\theta \log p(\mathbf{x};\theta)]^2 \leq \mathbb{V}_{p(\mathbf{x};\theta)}[f(\mathbf{x})]\mathbb{V}_{p(\mathbf{x};\theta)}[\nabla_\theta \log p(\mathbf{x};\theta)].$$

1. The score function gradient is a measure of covariance between the cost function and the score function.
2. (Cauchy-Schwartz inequality) the variance of the cost function is related to the magnitude and range of the gradient.

Overall cost: O(N(D + L)).

Score Function Estimator – Summary

- (Almost) any cost function can be used.
- The measure must be differentiable wrt. its parameters.
- We must be able to easily sample from the measure.
- It works for both continuous and discrete measures.
- It can be implemented with a single sample!
- Variance reduction is important.

Pathwise Gradient Estimator

- We can use structure of the measure to develop an estimator:

$$\hat{\mathbf{x}} \sim p(\mathbf{x};\boldsymbol{\theta}) \quad \equiv \quad \hat{\mathbf{x}} = g(\hat{\epsilon}, \boldsymbol{\theta}), \quad \hat{\epsilon} \sim p(\epsilon),$$

- These *sampling paths/processes* can be derived in a number of ways:
  - Change of variables: $p(\mathbf{x};\boldsymbol{\theta}) = p(\epsilon)\left|\nabla_{\epsilon} g(\epsilon;\boldsymbol{\theta})\right|^{-1}$.
  - Inversion methods: inverse CDF & uniform distribution.
  - Polar transformations: e.g. Box-Muller transform for sampling Gaussian random variables.
  - One-liners: $\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \equiv \hat{\mathbf{x}} = \boldsymbol{\mu} + \mathbf{L}\hat{\epsilon},\ \hat{\epsilon} \sim p(\epsilon),\ \mathbf{L}\mathbf{L}^{\top} = \Sigma$
- *Law of the Unconscious Statistician* (LOTUS):

$$\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right] = \mathbb{E}_{p(\epsilon)}\left[f(g(\epsilon;\boldsymbol{\theta}))\right]$$

Pathwise Gradient Estimator – Derivation

$$
\begin{aligned}
\boldsymbol{\eta} &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right] = \nabla_{\boldsymbol{\theta}} \int p(\mathbf{x};\boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x} \\
&= \nabla_{\boldsymbol{\theta}} \int p(\epsilon) f(g(\epsilon;\boldsymbol{\theta})) d\epsilon \\
&= \mathbb{E}_{p(\epsilon)}\left[\nabla_{\boldsymbol{\theta}} f(g(\epsilon;\boldsymbol{\theta}))\right].
\end{aligned}
$$

$$
\bar{\eta}_N = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} f(g(\hat{\epsilon}^{(n)};\boldsymbol{\theta})); \quad \hat{\epsilon}^{(n)} \sim p(\epsilon).
$$

Decoupling Sampling and Gradient Computation

$$
\begin{aligned}
\eta &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})} \left[f(\mathbf{x}) \right] \\
&= \mathbb{E}_{p(\epsilon)} \left[\nabla_{\boldsymbol{\theta}} f(\mathbf{x})|_{\mathbf{x}=g(\epsilon;\boldsymbol{\theta})} \right] \\
&= \int p(\epsilon) \nabla_{\mathbf{x}} f(\mathbf{x})|_{\mathbf{x}=g(\epsilon;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} g(\epsilon;\boldsymbol{\theta})\, d\epsilon \\
&= \int p(\mathbf{x};\boldsymbol{\theta}) \nabla_{\mathbf{x}} f(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \mathbf{x}\, d\mathbf{x} \\
&= \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})} \left[\nabla_{\mathbf{x}} f(\mathbf{x}) \nabla_{\boldsymbol{\theta}} \mathbf{x} \right]
\end{aligned}
$$

Pathwise Gradient Estimator – Bias & Variance

- Bias: we again interchanged order of integration and differentiation – cost function must be differentiable (i.e. no discontinuous cost functions allowed).
- Variance: is bounded by the squared Lipschitz constant of the cost function.
  - Bounds are independent of D.
  - As the cost becomes highly variable, the Lipschitz constant increases.

Pathwise Gradient Estimator – Computation

- For some discontinuous cost functions it is possible to smooth the function over the discontinuity and maintain the correctness of the gradient.
- Often multiple equivalent sampling paths. Not much theoretical motivation for choices – choose the simple one.
- Overall cost: $O(N(D + L))$.

Pathwise Gradient Estimator – Summary

- Only works for *differentiable* cost functions.
- Doesn't require an explicit measure – just base distribution and sampling path.
- Can be implemented using only a *single sample if needed*.
- May require *controlling the smoothness* of the function during learning to avoid large variance.
- May require variance reduction.

Measure-Valued Gradients

Weak derivative of $p(\mathbf{x};\theta)$

$$\nabla_{\theta_i} p(\mathbf{x};\theta) = c^+_{\theta_i} p^+(\mathbf{x};\theta) - c^-_{\theta_i} p^-(\mathbf{x};\theta),$$

Measure-Valued Gradients

Weak derivative of $p(\mathbf{x};\theta)$

$$\nabla_{\theta_i} p(\mathbf{x};\theta) = c_{\theta_i}\left(p^{+}(\mathbf{x};\theta) - p^{-}(\mathbf{x};\theta)\right).$$

- $\left(c_{\theta_i}, p^{+}, p^{-}\right)$
- Univariate definition is extended to the multivariate setting with a triple of vectors.
- Not unique, but always exists.
- Doesn't require p to be differentiable in its domain.

Measure-Valued Gradients

Weak derivative of $p(\mathbf{x};\theta)$

$$\nabla_{\theta_i} p(\mathbf{x};\boldsymbol{\theta}) = c_{\theta_i}\left(p^{+}(\mathbf{x};\boldsymbol{\theta}) - p^{-}(\mathbf{x};\boldsymbol{\theta})\right).$$

Distribution $p(x;\theta)$	Constant c_θ	Positive part $p^{+}(x)$	Negative part $p^{-}(x)$
Bernoulli(θ)	1	δ_1	δ_0
Poisson(θ)	1	$\mathcal{P}(\theta)+1$	$\mathcal{P}(\theta)$
Normal(θ,σ^2)	$1/\sigma\sqrt{2\pi}$	$\theta+\sigma\mathcal{W}(2,0.5)$	$\theta-\sigma\mathcal{W}(2,0.5)$
Normal(μ,θ^2)	$1/\theta$	$\mathcal{M}(\mu,\theta^2)$	$\mathcal{N}(\mu,\theta^2)$
Exponential(θ)	$1/\theta$	$\mathcal{E}(\theta)$	$\theta^{-1}\mathcal{E}r(2)$
Gamma(a,θ)	a/θ	$\mathcal{G}(a,\theta)$	$\mathcal{G}(a+1,\theta)$

Measure-Valued Gradients – Derivation

$$
\begin{aligned}
\eta_i &= \nabla_{\theta_i}\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right] = \nabla_{\theta_i}\int p(\mathbf{x};\boldsymbol{\theta})f(\mathbf{x})\mathrm{d}\mathbf{x} = \int \nabla_{\theta_i}p(\mathbf{x};\boldsymbol{\theta})f(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= c_{\theta_i}\left(\int f(\mathbf{x})p_i^{+}(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x} - \int f(\mathbf{x})p_i^{-}(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x}\right) \\
&= c_{\theta_i}\left(\mathbb{E}_{p_i^{+}(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right] - \mathbb{E}_{p_i^{-}(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right]\right) \\
\bar{\eta}_{i,N} &= \frac{c_{\theta_i}}{N}\left(\sum_{n=1}^{N} f(\dot{\mathbf{x}}^{(n)}) - \sum_{n=1}^{N} f(\ddot{\mathbf{x}}^{(n)})\right);\ \dot{\mathbf{x}}^{(n)} \sim p_i^{+}(\mathbf{x};\boldsymbol{\theta}),\ \ddot{\mathbf{x}}^{(n)} \sim p_i^{-}(\mathbf{x};\boldsymbol{\theta})
\end{aligned}
$$

Measure-Valued Gradients – Domination

- The score-function estimator used the dominated convergence theorem to establish correctness of the integral-derivative swap.
- The measure-valued estimator, allows the swap by definition:

$$\nabla_\theta \int f(x)p(x;\theta)\mathrm{d}x = c_\theta \left[\int f(x)p^+(x;\theta)\mathrm{d}x - \int f(x)p^-(x;\theta)\mathrm{d}x \right]$$

- No problems for $\mathcal{U}[0,\theta]$.

Measure-Valued Gradients – Bias & Variance

- Unbiased for bounded and continuous cost functions (by definition).
- Can also be shown to be unbiased for other types of cost functions.
- Variance:

$$\mathbb{V}_{p(\mathbf{x};\theta)}[\eta_N] = \mathbb{V}_{p^+(\mathbf{x};\theta)}[f(\mathbf{x})] + \mathbb{V}_{p^-(\mathbf{x};\theta)}[f(\mathbf{x})] - 2\mathrm{Cov}_{p^+p^-}[f(\mathbf{x}'), f(\mathbf{x})]$$

Measure-Valued Gradients – Computation

- Much more computationally expensive than either the score-function or pathwise estimators.
- Overall cost: O(2NDL) (vs. O(N(D + L))).
- Not applicable to very high-dimensional parameter spaces.
- BUT very low variance in most cases – trade-off.

Measure-Valued Gradients – Summary

- Can be used with any type of cost function, differentiable or not.
- Works for both discrete and continuous distributions.
- Computationally expensive in high-dimensional parameter spaces.
- Requires manual derivation of the decomposition.

Variance Reduction Techniques

- Large-samples
- Coupling
- Conditioning
- Control variates

Variance Reduction Techniques

- Large-samples
  - Easiest variance reduction technique.
  - Variance of our estimators shrinks as $\mathcal{O}(1/N)$.
- Coupling
- Conditioning
- Control variates

Variance Reduction Techniques

- Large-samples
- Coupling

$$\eta = \mathbb{E}_{p_1(\mathbf{x})}\left[f(\mathbf{x})\right] - \mathbb{E}_{p_2(\mathbf{x})}\left[f(\mathbf{x})\right]$$

$$\mathbb{V}_{p_{12}(\mathbf{x}_1,\mathbf{x}_2)}\left[\bar{\eta}_{\text{cpl}}\right] = \mathbb{V}_{p_1(\mathbf{x}_1)p_2(\mathbf{x}_2)}\left[\bar{\eta}_{\text{ind}}\right] - 2\text{Cov}_{p_{12}(\mathbf{x}_1,\mathbf{x}_2)}\left[f(\mathbf{x}_1), f(\mathbf{x}_2)\right]$$

- Conditioning
- Control variates

Variance Reduction Techniques

- Large-samples
- Coupling
- Conditioning
  - We Condition our estimators on a subset of dimensions and integrate out the remaining dimensions analytically.

$$\begin{aligned}\mathbb{V}_{p(\mathbf{x})}[f(\mathbf{x})] &= \mathbb{E}_{p(\mathbf{x}_{\mathcal{S}^c})}\left[\mathbb{V}_{p(\mathbf{x}_{\mathcal{S}})}\left[f(\mathbf{x})|\mathbf{x}_{\mathcal{S}^c}\right]\right] + \mathbb{V}_{p(\mathbf{x}_{\mathcal{S}^c})}[\mathbb{E}_{p(\mathbf{x}_{\mathcal{S}})}\left[f(\mathbf{x})|\mathbf{x}_{\mathcal{S}^c}\right]] \\ &\geq \mathbb{V}_{p(\mathbf{x}_{\mathcal{S}^c})}[\mathbb{E}_{p(\mathbf{x}_{\mathcal{S}})}\left[f(\mathbf{x})|\mathbf{x}_{\mathcal{S}^c}\right]]\end{aligned}$$

- Control variates

Variance Reduction Techniques – Control Variates

- Can be applied to any problem of the form $\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x})]$.
- Replace $f(\mathbf{x})$ with $\tilde{f}(\mathbf{x})$ whose expectation $\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\tilde{f}(\mathbf{x})\right]$ is the same, but whose variance is lower.

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - \beta(h(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[h(\mathbf{x})])$$

$$\bar{\eta}_N = \frac{1}{N}\sum_{n=1}^{N}\tilde{f}(\hat{\mathbf{x}}^{(n)}) = \bar{f} - \beta(\bar{h} - \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[h(\mathbf{x})]),$$

- The observed error $(h(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[h(\mathbf{x})])$ serves as a control in estimating $\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x})]$

Control Variates – Bias, Consistency & Variance

Unbiasedness:

$$\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\tilde{f}(\mathbf{x};\beta)\right] = \mathbb{E}\left[\bar{f} - \beta(\bar{h} - \mathbb{E}\left[h(\mathbf{x})\right])\right] = \mathbb{E}\left[\bar{f}\right] = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right]$$

Consistency: $$\lim_{n\to\infty} \frac{1}{N}\sum_{n=1}^{N}\tilde{f}(\hat{\mathbf{x}}^{(n)}) = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[\tilde{f}(\mathbf{x})\right] = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[f(\mathbf{x})\right]$$

Variance:

$$\frac{\mathbb{V}[\tilde{f}(\mathbf{x})]}{\mathbb{V}[f(\mathbf{x})]} = \frac{\mathbb{V}[f(\mathbf{x}) - \beta(h(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}\left[h(\mathbf{x})\right])]}{\mathbb{V}[f(\mathbf{x})]} = 1 - \mathrm{Corr}(f(\mathbf{x}), h(\mathbf{x}))^2$$

Closing Guidance I

- The pathwise estimator is a good default for continuous functions and measures that are continuous in the domain.
- If the cost function is non-differentiable or black-box then the score-function or the measure-valued gradients will work.
- The score-function should always be implemented with some kind of variance reduction.
- For the score-function estimator, the dynamic range of the cost function and its variance should be monitored, and ways found to keep its value bounded within a reasonable range.
- For all estimators, track the variance of the gradients and address problems by using a larger number of samples, a lower learning rate, or clipping the gradient values.

Closing Guidance II

- The measure-valued gradient should be used with a coupling method for variance reduction
- With several unbiased gradient estimators, a convex combination might have lower variance.
- For measures discrete on their domain then use the score-function or measure-valued gradient.
- In all cases, implement a broad set of tests to verify unbiasedness of the gradient estimator.

Thanks for listening!