Approximating Gradients of Expectations and It's Applications

Deniz Koyuncu

December 4, 2019

Overview

- First Half: Approximating Gradients of Expectations
  - A Review, Approximating Expectations
  - A General Objective Function
  - VI Objective Function
  - Score Functions
  - Pathwise Gradient Estimators
- Second Half: Application to VAE
  - Models with Unobserved Variables
  - EM Algorithm
  - VAE Model Description
  - Connection to EM

Approximating Expectations

$$\mathrm{E}[\mathrm{X}] = \int P(\mathrm{X} = x) x dx$$

- How can we empirically calculate the expectation?
- Assume $\mathrm{X}_1, \ldots, \mathrm{X}_N$ are i.i.d samples of X.

$$\begin{aligned} \hat{\mu} &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{X}_i \\ \mathrm{E}[\hat{\mu}] &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[\mathrm{X}_i] = \mathrm{E}[\mathrm{X}] \\ \mathrm{Var}(\hat{\mu}) &= \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(\mathrm{X}_i) = \frac{\mathrm{Var}(\mathrm{X})}{N} \end{aligned} \tag{1}$$

- $\hat{\mu}$ is unbiased and notice that for finite variance $\lim_{N \to \infty} \mathrm{Var}(\hat{\mu}) = 0$

Expectation of a Function of a R.V

$$\mathrm{E}[g(\mathrm{X})] = \int P(\mathrm{X} = x) g(x) dx$$

- Similarly, assume $\mathrm{X}_1, \ldots, \mathrm{X}_N$ are i.i.d samples of X.

$$\begin{aligned}
\hat{\mu_g} &= \frac{1}{N} \sum_{i=1}^{N} g(\mathrm{X}_i) \\
\mathrm{E}[\hat{\mu_g}] &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[g(\mathrm{X}_i)] = \mathrm{E}[g(\mathrm{X})] \\
\mathrm{Var}(\hat{\mu_g}) &= \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(g(\mathrm{X}_i)) = \frac{\mathrm{Var}(g(\mathrm{X}))}{N}
\end{aligned} \tag{2}$$

- Again, $\hat{\mu_g}$ is unbiased and for finite variance $\lim_{N \to \infty} \mathrm{Var}(\hat{\mu_g}) = 0$

A Specific Form of Objective Function

- Following [1]'s derivation,
- In our discussion $\mathbf{X}$ is assumed to be continues random vector
- Today, we will be talking about the objective functions in the following form

$$\mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] = \int P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x}$$

$$\begin{aligned} \boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] &= \nabla_{\boldsymbol{\theta}} \int P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} \\ &= \int \nabla_{\boldsymbol{\theta}} P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{3}$$

- Notice that Eq. 3, is no longer an expectation and the integral may not have a closed form
- We're interested in somehow turning it into an expectation so that we can empirically estimate it

VI Objective Function

- VI tries to approximate $P(\mathbf{X}_h \mid \mathbf{X}_o = \mathbf{x}_o)$, with another distribution when exact calculation is not an option.
- One distance measure is the following KL divergence:

$$q^* = \arg\min_q KL(q(\mathbf{X}_h) \parallel P(\mathbf{X}_h \mid \mathbf{X}_o = \mathbf{x}_o)) \tag{4}$$

- It's minimum when they're equal.
- The generally intractable term $P(\mathbf{X}_o = \mathbf{x}_o)$ is a constant for q.

Gradient Based VI

- Previously we were optimizing in the function space q,
- In order to use gradient based optimization, we parameterize q as $q(\mathbf{X}_h; \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ are the parameters,
- Therefore, the optimization function becomes,

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} KL(q(\mathbf{X}_h; \boldsymbol{\lambda}) \parallel P(\mathbf{X}_h \mid \mathbf{X}_o = \mathbf{x}_o)) \tag{5}$$

- One can take the gradient as follows

Calculating the Gradients

$$
\begin{aligned}
&\nabla_{\boldsymbol{\lambda}} KL(q(\mathbf{X}_h; \boldsymbol{\lambda}) \parallel P(\mathbf{X}_h \mid \mathbf{X}_o = \mathbf{x}_o)) \\
&\quad = \nabla_{\boldsymbol{\lambda}} \int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) \log \frac{q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda})}{P(\mathbf{X}_h = \mathbf{x}_h, \mathbf{X}_o = \mathbf{x}_o)} d\mathbf{x}_h \\
&\quad = -\nabla_{\boldsymbol{\lambda}} \int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) \log P(\mathbf{X}_h = \mathbf{x}_h, \mathbf{X}_o = \mathbf{x}_o) d\mathbf{x}_h \\
&\quad + \nabla_{\boldsymbol{\lambda}} \int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) \log q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) d\mathbf{x}_h
\end{aligned} \tag{6}
$$

- Observe that the second term is the - entropy of q,

$$
H_q(\mathbf{X}; \boldsymbol{\lambda}) = -\int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) \log q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) d\mathbf{x}_h
$$

- If q is a exponential family than entropy has an analytical form.

The First Term

$$\begin{aligned}\nabla_{\boldsymbol{\lambda}} \int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) \log P(\mathbf{X}_h = \mathbf{x}_h, \mathbf{X}_o = \mathbf{x}_o) d\mathbf{x}_h \\ = \nabla_{\boldsymbol{\lambda}} \int q(\mathbf{X}_h = \mathbf{x}_h; \boldsymbol{\lambda}) g(\mathbf{x}_h) d\mathbf{x}_h \\ = \nabla_{\boldsymbol{\lambda}} \mathrm{E}_{q(\mathbf{x}_h; \boldsymbol{\lambda})}[g(\mathbf{X}_h)]\end{aligned} \tag{7}$$

- It's in the same form with $\boldsymbol{\eta}$.

2 Approaches

- We're going to be talking about two approaches to approximate the gradient $\boldsymbol{\eta}$,
    1. Score Functions
    2. Pathwise Gradient Estimators

Score Functions

- One way of turning Eq. 3 into an expectation is as follows

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] &= \int \nabla_{\boldsymbol{\theta}} P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} \\ &= \int \frac{P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta})}{P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x} \\ &= \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[\frac{\nabla_{\boldsymbol{\theta}} P(\mathbf{X}; \boldsymbol{\theta})}{P(\mathbf{X}; \boldsymbol{\theta})} g(\mathbf{X})]\end{aligned} \tag{8}$$

- Now, since it's a expectation we can do the empirical estimation.

Empirical Expectation

- Using $\nabla \log g(\mathbf{x}) = \dfrac{\nabla g(\mathbf{x})}{g(\mathbf{x})}$

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] &= \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log P(\mathbf{X};\boldsymbol{\theta}) g(\mathbf{X})] \\ &= \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[\vec{h}(\mathbf{X})] \end{aligned} \tag{9}$$

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \frac{1}{N} \sum_{s=1}^{N} \vec{h}(\mathbf{x}^{(s)}) \\ &= \frac{1}{N} \sum_{s=1}^{N} \nabla_{\boldsymbol{\theta}} \log P(\mathbf{X} = \mathbf{x}^{(s)};\boldsymbol{\theta}) g(\mathbf{x}^{(s)}), \end{aligned} \tag{10}$$

$$\text{where} \quad \mathbf{x}^{(s)} \sim P(\mathbf{X};\boldsymbol{\theta})$$

Comments

- $\hat{\eta}$ is an unbiased estimator of η.
- As N increases, variance of gradient of each parameter decrease.
- There are interesting interpretations of the variance of score function estimator one can refer to [1].
- There are different approaches to reduce the variance of the score function estimator.

Pathwise Gradient Estimators

- In cases where the random vector **X** can be written as a function of another random vector ϵ,

$$\boldsymbol{X} = t(\epsilon; \boldsymbol{\theta})$$

- And one can sample from ϵ, we can use pathwise gradient estimators.
- One example is MVN [1],

$$\begin{aligned} \mathbf{X} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \epsilon &\sim N(\mathbf{0}, \mathbf{I}) \\ \mathbf{X} &= t(\epsilon; \boldsymbol{\theta}) = \boldsymbol{\mu} + \mathbf{L}\epsilon, \quad \text{where } \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma} \end{aligned} \tag{11}$$

Derivations

- The following approach can be used.

$$\begin{aligned} \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] &= \mathrm{E}_{P(\boldsymbol{\epsilon})}[g(t(\boldsymbol{\epsilon};\boldsymbol{\theta}))] \\ &= \int P(\boldsymbol{\epsilon}=\varepsilon)g(t(\varepsilon;\boldsymbol{\theta}))d\varepsilon \end{aligned} \tag{12}$$

- Notice that the distribution the expectation is with respect to namely $P(\boldsymbol{\epsilon})$, is independent of $\boldsymbol{\theta}$ this time.

Frame Title

- Taking the gradient,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathrm{E}_{P(\mathbf{X};\boldsymbol{\theta})}[g(\mathbf{X})] &= \nabla_{\boldsymbol{\theta}} \int P(\boldsymbol{\epsilon} = \varepsilon) g(t(\varepsilon; \boldsymbol{\theta})) d\varepsilon \\ &= \int P(\boldsymbol{\epsilon} = \varepsilon) \nabla_{\boldsymbol{\theta}} g(t(\varepsilon; \boldsymbol{\theta})) d\varepsilon \\ &= \mathrm{E}_{P(\boldsymbol{\epsilon})}[\vec{p}(\boldsymbol{\epsilon})]\end{aligned} \tag{13}$$

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= \frac{1}{N} \sum_{s=1}^{N} \vec{p}(\varepsilon^{(s)}) \\ &= \frac{1}{N} \sum_{s=1}^{N} g(t(\varepsilon^{(s)}; \boldsymbol{\theta})) \\ &\text{where} \quad \varepsilon^{(s)} \sim P(\boldsymbol{\epsilon})\end{aligned} \tag{14}$$

Comments

- $\hat{\eta}$ is an unbiased estimator of η and the variance goes to zero for each parameter.
- A more interesting remark is the comparison of two gradient estimators we have covered.
  - In [1], one remark made is the bound on the variance of the pathwise gradient estimator doesn't depend on the number of parameters.
  - The bound of variance of score function estimator depends on the number of parameters.
  - This doesn't imply that pathwise approach always have lower variance. [1]
- A meaningful question to ask can be which distributions can be written in terms other simpler distributions.

Shakir Mohamed et al. *Monte Carlo Gradient Estimation in Machine Learning*. 2019. arXiv: 1906.10652 [stat.ML].