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

\[ E[X] = \int P(X = x) \, dx \]

- How can we empirically calculate the expectation?
- Assume \( X_1, \ldots, X_N \) are i.i.d samples of \( X \).

\[ \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i \]

\[ E[\hat{\mu}] = \frac{1}{N} \sum_{i=1}^{N} E[X_i] = E[X] \quad (1) \]

\[ \text{Var}(\hat{\mu}) = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}(X_i) = \frac{\text{Var}(X)}{N} \]

- \( \hat{\mu} \) is unbiased and notice that for finite variance \( \lim_{N \to \infty} \text{Var}(\hat{\mu}) = 0 \)
Expectation of a Function of a R.V

\[ E[g(X)] = \int P(X = x)g(x)dx \]

- Similarly, assume \( X_1, \ldots, X_N \) are i.i.d samples of \( X \).

\[ \hat{\mu}_g = \frac{1}{N} \sum_{i=1}^{N} g(X_i) \]

\[ E[\hat{\mu}_g] = \frac{1}{N} \sum_{i=1}^{N} E[g(X_i)] = E[g(X)] \tag{2} \]

\[ \text{Var}(\hat{\mu}_g) = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}(g(X_i)) = \frac{\text{Var}(g(X))}{N} \]

- Again, \( \hat{\mu}_g \) is unbiased and for finite variance \( \lim_{N \to \infty} \text{Var}(\hat{\mu}_g) = 0 \)
A Specific Form of Objective Function

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

$$E_{P(X;\theta)}[g(X)] = \int P(X = x; \theta)g(x)\,dx$$

$$\eta = \nabla_\theta E_{P(X;\theta)}[g(X)] = \nabla_\theta \int P(X = x; \theta)g(x)\,dx$$

$$= \int \nabla_\theta P(X = x; \theta)g(x)\,dx$$  \hspace{1cm} (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 $P(\mathbf{X}_h \mid \mathbf{X}_o = x_o)$, with another distribution when exact calculation is not an option.
- One distance measure is the following KL divergence:

$$q^* = \arg\min_{q} KL(q(\mathbf{X}_h) \parallel P(\mathbf{X}_h \mid \mathbf{X}_o = x_o))$$  \hspace{1cm} (4)

- It’s minimum when they’re equal.
- The generally intractable term $P(\mathbf{X}_o = x_o)$ is a constant for $q$. 
Previously we were optimizing in the function space $q$,

In order to use gradient based optimization, we parameterize $q$ as

$q(\mathbf{X}_h; \lambda)$ where $\lambda$ are the parameters,

Therefore, the optimization function becomes,

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

(5)

One can take the gradient as follows
Calculating the Gradients

\[ \nabla_\lambda KL(q(X_h; \lambda) \parallel P(X_h \mid X_o = x_o)) \]

\[ = \nabla_\lambda \int q(X_h = x_h; \lambda) \log \frac{q(X_h = x_h; \lambda)}{P(X_h = x_h, X_o = x_o)} dX_h \]

\[ = -\nabla_\lambda \int q(X_h = x_h; \lambda) \log P(X_h = x_h, X_o = x_o) dX_h \]

\[ + \nabla_\lambda \int q(X_h = x_h; \lambda) \log q(X_h = x_h; \lambda) dX_h \]

(6)

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

\[ H_q(X; \lambda) = -\int q(X_h = x_h; \lambda) \log q(X_h = x_h; \lambda) dX_h \]

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

\[ \nabla_{\lambda} \int q(X_h = x_h; \lambda) \log P(X_h = x_h, X_o = x_o) \, dx_h \]

\[ = \nabla_{\lambda} \int q(X_h = x_h; \lambda) g(x_h) \, dx_h \]

\[ = \nabla_{\lambda} \mathbb{E}_{q(X_h; \lambda)}[g(X_h)] \]

- It’s in the same form with \( \eta \).
2 Approaches

We’re going to be talking about two approaches to approximate the gradient $\eta$,

1. Score Functions
2. Pathwise Gradient Estimators
One way of turning Eq. 3 into an expectation is as follows

\[
\nabla_\theta E_P(x; \theta)[g(x)] = \int \nabla_\theta P(x = x; \theta) g(x) dx \\
= \int \frac{P(x = x; \theta)}{P(x = x; \theta)} \nabla_\theta P(x = x; \theta) g(x) dx \\
= E_P(x; \theta)[\frac{\nabla_\theta P(x; \theta)}{P(x; \theta)} g(x)]
\]

Now, since it’s a expectation we can do the empirical estimation.
Empirical Expectation

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

\begin{align*}
\nabla \theta E_P(x; \theta)[g(X)] &= E_P(x; \theta)[\nabla \theta \log P(X; \theta)g(X)] \\
&= E_P(x; \theta)[\hat{h}(X)]
\end{align*}

(9)

\begin{align*}
\hat{\eta} &= \frac{1}{N} \sum_{s=1}^{N} \text{h}(x^{(s)}) \\
&= \frac{1}{N} \sum_{s=1}^{N} \nabla \theta \log P(X = x^{(s)}; \theta)g(x^{(s)}), \quad \text{(10)}
\end{align*}

where $x^{(s)} \sim P(X; \theta)$
• $\hat{\eta}$ is an unbiased estimator of $\eta$.
• 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 $\epsilon$,
  \[ X = t(\epsilon; \theta) \]
- And one can sample from $\epsilon$, we can use pathwise gradient estimators.
- One example is MVN [1],
  \[ X \sim N(\mu, \Sigma) \]
  \[ \epsilon \sim N(0, I) \]
  \[ X = t(\epsilon; \theta) = \mu + L\epsilon, \quad \text{where } LL^T = \Sigma \]
The following approach can be used.

\[
E_{P(x;\theta)}[g(X)] = E_{P(\epsilon)}[g(t(\epsilon; \theta))]
= \int P(\epsilon = \epsilon)g(t(\epsilon; \theta))d\epsilon
\]

Notice that the distribution the expectation is with respect to namely \(P(\epsilon)\), is independent of \(\theta\) this time.
Taking the gradient,

\[ \nabla_\theta E_{P(X;\theta)}[g(X)] = \nabla_\theta \int P(\epsilon = \epsilon) g(t(\epsilon; \theta)) d\epsilon \]

\[ = \int P(\epsilon = \epsilon) \nabla_\theta g(t(\epsilon; \theta)) d\epsilon \]

\[ = E_{P(\epsilon)}[\bar{p}(\epsilon)] \]

\[ \hat{\eta} = \frac{1}{N} \sum_{s=1}^{N} \bar{p}(\epsilon^{(s)}) \]

\[ = \frac{1}{N} \sum_{s=1}^{N} g(t(\epsilon^{(s)}; \theta)) \]

where \( \epsilon^{(s)} \sim P(\epsilon) \)
- $\hat{\eta}$ is an unbiased estimator of $\eta$ 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.