Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee   Hangyeol Yu   Hongseok Yang
KAIST

Published at NeurIPS 2018

Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee   Hangyeol Yu   Hongseok Yang
KAIST

Published at NeurIPS 2018

Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee Hangyeol Yu Hongseok Yang
KAIST

Published at NeurIPS 2018

Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee   Hangyeol Yu   Hongseok Yang
KAIST

Published at NeurIPS 2018

Backgrounds

Posterior inference

- Latent variable $z \in \mathbb{R}^n$.
- Observed variable $x \in \mathbb{R}^m$.
- Joint density $p(x,z)$.
- Want to infer posterior $p(z|x^0)$ given a particular value x^0 of x.

Variational inference

1. Fix a family of variational distr. $\{q_\theta(z)\}_\theta$.
2. Find $q_\theta(z)$ that approximates $p(z|x^0)$ well.

Variational inference

differentiable & easy-to-sample

1. Fix a family of variational distr. $\{q_\theta(z)\}_\theta$.
2. Find $q_\theta(z)$ that approximates $p(z|x^0)$ well.

Variational inference

differentiable & easy-to-sample

1. Fix a family of variational distr. $\{q_\theta(z)\}_\theta$.
2. Find $q_\theta(z)$ that approximates $p(z|x^0)$ well.

Typically, by solving

$$\text{argmax}_\theta(\text{ELBO}_\theta)$$

where $\text{ELBO}_\theta = \mathbb{E}_{q_\theta(z)}[\ \log(\ p(x^0,z)/q_\theta(z)\ )\ ]$.

Variational inference

differentiable & easy-to-sample

1. Fix a family of variational distr. $\{q_\theta(z)\}_\theta$.
2. Find $q_\theta(z)$ that approximates $p(z|x^0)$ well.

Typically, by solving

$$\text{argmax}_\theta(\text{ELBO}_\theta)$$

where $\text{ELBO}_\theta = \mathbb{E}_{q\theta(z)}[\quad ..\ z\ ..\ z\ ..\quad].$

Gradient ascent

$$\theta_{n+1} = \theta_n + \eta \times \nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}$$

Gradient ascent

$$\theta_{n+1} = \theta_n + \eta \times \nabla_\theta \mathrm{ELBO}_{\theta=\theta n}$$

- Difficult to compute $\nabla_\theta \mathrm{ELBO}_\theta$.

Gradient ascent

$$\theta_{n+1} = \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

- Difficult to compute $\nabla_\theta \mathrm{ELBO}_\theta$.
- Use an estimated gradient instead.

Reparameterization estimator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.
- Derived from the equation:

$$\nabla_\theta \text{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\nabla_\theta(..\, f_\theta(\varepsilon)\, ..\, f_\theta(\varepsilon)\, ..)]$$

Reparameterization estimator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.

- Derived from the equation:

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\ \nabla_\theta(..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..)\ ]$$

$\nabla_\theta ELBO_\theta =$

imator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.

- Derived from the equation:

$$\nabla_\theta ELBO_\theta = \mathbb{E}_{q(\varepsilon)}[\ \nabla_\theta(..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..)\ ]$$

$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q\theta(z)}[.. z .. z ..]$

imator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.

- Derived from the equation:

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\, \nabla_\theta(..\, f_\theta(\varepsilon)\, ..\, f_\theta(\varepsilon)\, ..)\,]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q_\theta(z)}[..\ z\ ..\ z\ ..]$$

$$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..]$$

imator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.

- Derived from the equation:

$$\nabla_\theta \text{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\ \nabla_\theta(..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..)\ ]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q\theta(z)}[..\ z\ ..\ z\ ..]$$
$$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..]$$
$$= \mathbb{E}_{q(\varepsilon)}[\nabla_\theta(..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..)]$$

imator

- Works if $p(x^0,z)$ is differentiable wrt. z.
- Need distr. $q(\varepsilon)$ & smooth function $f_\theta(\varepsilon)$ s.t.

  $f_\theta(\varepsilon)$ for $\varepsilon \sim q(\varepsilon)$ has the distr. $q_\theta(z)$.

- Derived from the equation:

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\ \nabla_\theta(..\ f_\theta(\varepsilon)\ ..\ f_\theta(\varepsilon)\ ..)\ ]$$

Non-differentiable models
from probabilistic programming

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$

$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$

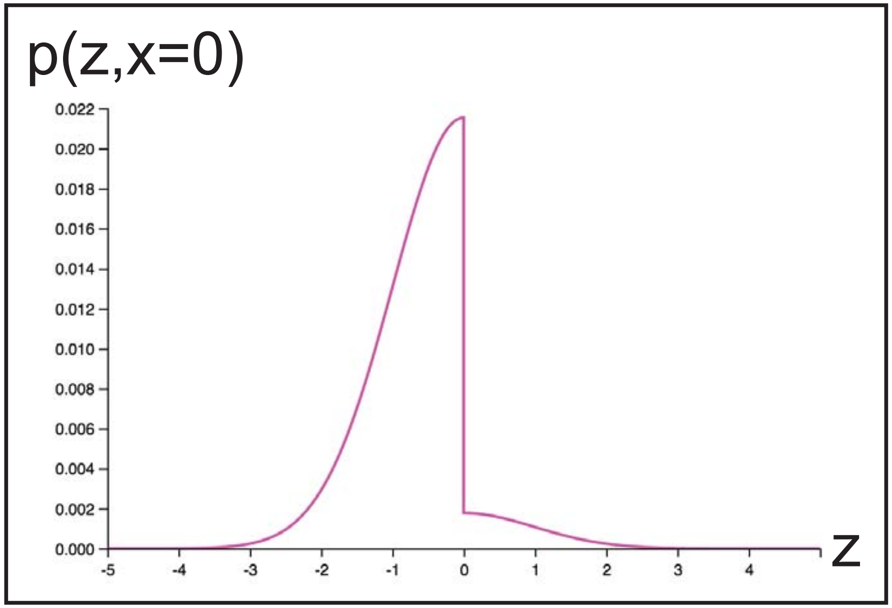
$$p(z,x=0) = [z>0]r_1(z) + [z \leq 0]r_2(z)$$

```
(let
  [z (sample (normal 0 1))]
  (if (> z 0)
    (observe (normal 3 1) 0)
    (observe (normal -2 1) 0))
  z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$



```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

$\approx$

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$$
$$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$$
$$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

How to find a good θ?

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?

By gradient ascent on $ELBO_\theta$.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \nabla_\theta \, ELBO_{\theta=\theta_n}$$

```
(let
   [z (sample (normal 0 1))]
   (if (> z 0)
      (observe (normal 3 1) 0)
      (observe (normal -2 1) 0))
   z)
```

≈

```
(let
   [ε (sample (normal 0 1))
     z (+ ε θ)]
   z)
```

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$\nabla_\theta \mathrm{ELBO}_\theta$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta n}}$$

$\nabla_\theta \mathrm{ELBO}_\theta$

$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\log(r_2(\varepsilon+\theta))]$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

How to find a good θ?

By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$\nabla_\theta \text{ELBO}_\theta$

$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\log(r_2(\varepsilon+\theta))]$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

How to find a good θ?

By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \text{ELBO}_{\theta=\theta_n}}$$

$$\nabla_\theta \mathrm{ELBO}_\theta$$
$$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\log(r_2(\varepsilon+\theta))]$$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta n}}$$

$$\nabla_\theta \mathrm{ELBO}_\theta$$
$$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\log(r_2(\varepsilon+\theta))]$$
$$= \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\nabla_\theta\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\nabla_\theta\log(r_2(\varepsilon+\theta))]$$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$\nabla_\theta \text{ELBO}_\theta$

$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon \le -\theta]\log(r_2(\varepsilon+\theta))]$

$= \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\nabla_\theta \log(r_1(\varepsilon+\theta)) + [\varepsilon \le -\theta]\nabla_\theta \log(r_2(\varepsilon+\theta))]$

$= \mathbb{E}_{q(\varepsilon)}[\ -\theta-\varepsilon\ ] = -\theta$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\le 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

How to find a good θ?

By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \text{ELBO}_{\theta=\theta n}}$$

$$
\begin{aligned}
&\nabla_\theta \mathrm{ELBO}_\theta \\
&= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\log(r_2(\varepsilon+\theta))] \\
&= \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\nabla_\theta\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\nabla_\theta\log(r_2(\varepsilon+\theta))] \\
&= \mathbb{E}_{q(\varepsilon)}[\,-\theta-\varepsilon\,] = -\theta
\end{aligned}
$$

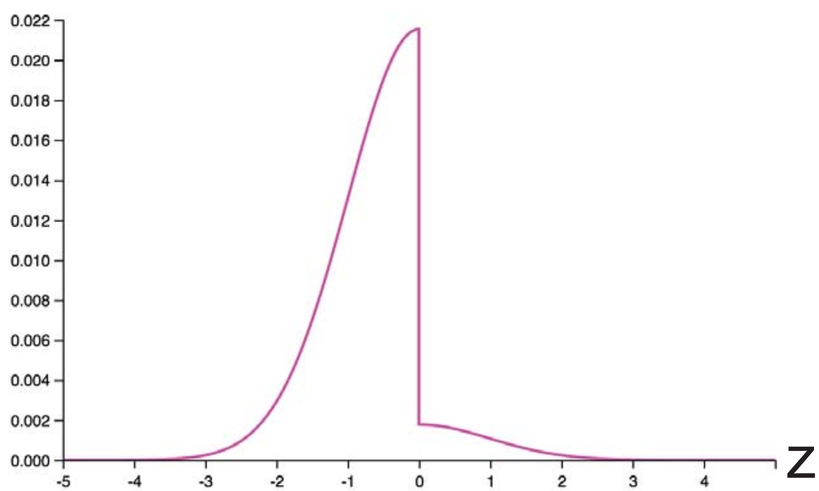
$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$$\nabla_\theta \mathrm{ELBO}_\theta$$
$$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\log(r_2(\varepsilon+\theta))]$$
$$= \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\nabla_\theta \log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\nabla_\theta \log(r_2(\varepsilon+\theta))]$$
$$= \mathbb{E}_{q(\varepsilon)}[\, -\theta-\varepsilon \,] = -\theta$$



$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

BO$_\theta$.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$$
\begin{aligned}
&\nabla_\theta \mathrm{ELBO}_\theta \\
&= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\log(r_2(\varepsilon+\theta))] \\
&\neq \mathbb{E}_{q(\varepsilon)}[[\varepsilon>-\theta]\nabla_\theta\log(r_1(\varepsilon+\theta)) + [\varepsilon\leq-\theta]\nabla_\theta\log(r_2(\varepsilon+\theta))] \\
&= \mathbb{E}_{q(\varepsilon)}[\,-\theta-\varepsilon\,] = -\theta
\end{aligned}
$$

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$

$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$

$p(z,x=0) = [z>0]r_1(z) + [z\leq0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$

$z = \varepsilon+\theta$

How to find a good θ?

By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

$\nabla_\theta \mathrm{ELBO}_\theta$
$= \nabla_\theta \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\log(r_2(\varepsilon+\theta))]$
$= \mathbb{E}_{q(\varepsilon)}[[\varepsilon > -\theta]\nabla_\theta \log(r_1(\varepsilon+\theta)) + [\varepsilon \leq -\theta]\nabla_\theta \log(r_2(\varepsilon+\theta))]$
$+$ Correction Term

$r_1(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|3,1)$
$r_2(z) = \mathcal{N}(z|0,1)\mathcal{N}(x=0|-2,1)$
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

$q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$
$z = \varepsilon+\theta$

How to find a good θ?
By gradient ascent on ELBO_θ.

$$\theta_{n+1} \leftarrow \theta_n + \eta \times \widehat{\nabla_\theta \mathrm{ELBO}_{\theta=\theta_n}}$$

Why doesn't it work?

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x = \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

- Careful when exchanging gradient and integration.

Why doesn't it work?

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x \neq \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

- Careful when exchanging gradient and integration.
- May fail unexpectedly.

Why doesn't it work?

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x = \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

+ CorrectionTerm

- Careful when exchanging gradient and integration.
- May fail unexpectedly.
- May hold unexpectedly, but with correction.

Our results formally

Non-differentiable models

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

Non-differentiable models

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

- $z \in \mathbb{R}^n$.

Non-differentiable models

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

- $z \in \mathbb{R}^n$.
- $\{A_1, \cdots, A_K\}$ forms a **partition** of $\mathbb{R}^n$.
- r_k is **differentiable.**

Non-differentiable models

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

- $z \in \mathbb{R}^n$.
- $\{A_1, \cdots, A_K\}$ forms a **partition** of $\mathbb{R}^n$.
- r_k is **differentiable**.
- ∂A_k has Lebesgue **measure zero**.

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k [f_\theta(\epsilon) \in A_k] \cdot H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k [f_\theta(\epsilon) \in A_k] \cdot H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

Wishful thinking

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\neq \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

Correction

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$$+ \sum_k \text{ surface integral over } \partial f_\theta^{-1}(A_k)$$

Correction

$$\int_{B_{k,\theta}} \cdots \; q(\epsilon)d\epsilon$$

$$[z \in A_k] \cdot r_k(z)$$

$$\mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$$+ \sum_k \text{ surface integral over } \partial f_\theta^{-1}(A_k)$$

Correction

$$\nabla_\theta \int_{B_{k,\theta}} \cdots \; q(\epsilon)d\epsilon \qquad [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

$$\nabla_\theta \int_{B_{k,\theta}} \cdots\ q(\epsilon) d\epsilon$$

$$= \int_{B_{k,\theta}} \nabla_\theta \big(\cdots\ q(\epsilon)\big) d\epsilon + \int_{\partial B_{k,\theta}} (\cdots) \cdot d\boldsymbol{\Sigma}$$

$$\mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

$$\nabla_\theta \int_{B_{k,\theta}} \cdots\ q(\epsilon)d\epsilon$$

$$= \int_{B_{k,\theta}} \nabla_\theta(\cdots\ q(\epsilon))d\epsilon + \int_{\partial B_{k,\theta}} (\cdots) \cdot d\mathbf{\Sigma}$$

$$\mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}[\textstyle\sum_k[\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon,\theta)]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\textstyle\sum_k[\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon,\theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\textstyle\sum_k[\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon,\theta)\right]$$

$$+ \textstyle\sum_k \text{ surface integral over } \partial f_\theta^{-1}(A_k)$$

Correction

$$\nabla_\theta \int_{B_{k,\theta}} \cdots \, q(\epsilon) d\epsilon$$

$$= \int_{B_{k,\theta}} \nabla_\theta \big(\cdots \, q(\epsilon)\big) d\epsilon + \int_{\partial B_{k,\theta}} (\cdots) \cdot d\boldsymbol{\Sigma}$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\textstyle\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\textstyle\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$$+ \textstyle\sum_k \text{ surface integral over } \partial f_\theta^{-1}(A_k)$$

Correction

$$\nabla_\theta \int_{B_{k,\theta}} \cdots\ q(\epsilon) d\epsilon$$

$$= \int_{B_{k,\theta}} \nabla_\theta \big(\cdots\ q(\epsilon)\big) d\epsilon + \int_{\partial B_{k,\theta}} (\cdots) \cdot d\boldsymbol{\Sigma}$$

$$\mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

$$p(x^0, z) = \sum_k [z \in A_k] \cdot r_k(z)$$

$$\text{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \text{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+\sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

Accounts for the impact of **moving** the boundaries.

$$\mathrm{ELBO}_\theta = \quad \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Correction

Accounts for the impact of **moving** the boundaries.

Can be estimated by **manifold sampling**.

$$\mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$\nabla_\theta \mathrm{ELBO}_\theta = \nabla_\theta \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot H_k(\epsilon, \theta)\right]$$

$$= \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$$+ \sum_k \text{surface integral over } \partial f_\theta^{-1}(A_k)$$

Two ingredients

- Differentiation under moving domain:

$$\nabla_\theta \int_{B_\theta} g(\epsilon,\theta)d\epsilon = \int_{B_\theta} \big(\nabla_\theta g + \nabla_\epsilon \cdot (g\mathbf{V})\big)(\epsilon,\theta)d\epsilon$$

Two ingredients

- Differentiation under moving domain:

$$\nabla_\theta \int_{B_\theta} g(\epsilon,\theta)d\epsilon = \int_{B_\theta} \left(\nabla_\theta g + \nabla_\epsilon \cdot (g\mathbf{V})\right)(\epsilon,\theta)d\epsilon$$

- Divergence theorem:

$$\int_B (\nabla \cdot \mathbf{G})\mathrm{d}V = \int_{\partial B} \mathbf{G} \cdot \mathrm{d}\mathbf{\Sigma}$$

Two ingredients

$$\nabla_\theta \int_{B_\theta} g(\epsilon,\theta)d\epsilon = \int_{B_\theta} \big(\nabla_\theta g + \nabla_\epsilon \cdot (g\mathbf{V})\big)(\epsilon,\theta)d\epsilon$$

$$= \int_{B_\theta} \nabla_\theta g(\epsilon,\theta)d\epsilon + \int_{\partial B_\theta} (g\mathbf{V})(\epsilon,\theta)\cdot \mathrm{d}\mathbf{\Sigma}$$

Correction term

Surface integral over $\partial f_\theta^{-1}(A_k)$

$$= \int_{\partial f_\theta^{-1}(A_k)} \big(q(\epsilon) H_k(\epsilon, \theta) \mathbf{V}(\epsilon, \theta)\big) \cdot \mathrm{d}\mathbf{\Sigma}$$

- $\mathbf{V}(\epsilon, \theta)_{ij} = \left(\frac{\partial f_\theta}{\partial \theta_i}\right)_j$
- $\mathbf{\Sigma}$ is a normal vector of $\partial f_\theta^{-1}(A_k)$.

Correction term

Surface integral over $\partial f_\theta^{-1}(A_k)$

$$= \int_{\partial f_\theta^{-1}(A_k)} \left(q(\epsilon)H_k(\epsilon,\theta)\mathbf{V}(\epsilon,\theta)\right) \cdot \mathrm{d}\boldsymbol{\Sigma}$$

Requires manifold sampling

Hard to estimate in general cases

Correction term

Surface integral over $\partial f_{\theta}^{-1}(A_k)$

$$= \int_{\partial f_{\theta}^{-1}(A_k)} \big(q(\epsilon) H_k(\epsilon, \theta) \mathbf{V}(\epsilon, \theta)\big) \cdot \mathrm{d}\boldsymbol{\Sigma}$$

- Easy to estimate if $\partial f_{\theta}^{-1}(A_k)$ is a **hyperplane**.

Correction term

Surface integral over $\partial f_\theta^{-1}(A_k)$

$$= \int_{\partial f_\theta^{-1}(A_k)} \big(q(\epsilon) H_k(\epsilon, \theta) \mathbf{V}(\epsilon, \theta)\big) \cdot \mathrm{d}\boldsymbol{\Sigma}$$

- Easy to estimate if $\partial f_\theta^{-1}(A_k)$ is a **hyperplane**.
- Assume the branch condition of each if-statement is **linear** in z.

Subsampling k

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k [\epsilon \in f_\theta^{-1}(A_k)] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$+ \sum_k$ surface integral over $\partial f_\theta^{-1}(A_k)$

Subsampling k

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\epsilon)}\left[\sum_k \left[\epsilon \in f_\theta^{-1}(A_k)\right] \cdot \nabla_\theta H_k(\epsilon, \theta)\right]$$

$$+ \sum_k \text{ surface integral over } \partial f_\theta^{-1}(A_k)$$

- For computational efficiency,
  we **subsample** k surface integrals.

Experiments

Implementation

- Implemented a **black-box variational inference** engine for a simple probabilistic programming language
- Supports `sample`, `observe`, `if`, ...
- Written in Python, using **`autograd`** package.

Benchmarks

`textmsg`

- Models #’s of per-day **SNS msg’s**, where SNS-usage pattern changes on some day.
- **Non-differentiable** part: the day of change in SNS-usage pattern.
- Given #’s of per-day SNS msg’s over 2 months, infer **the day** when the pattern changes.

Benchmarks

`temperature`

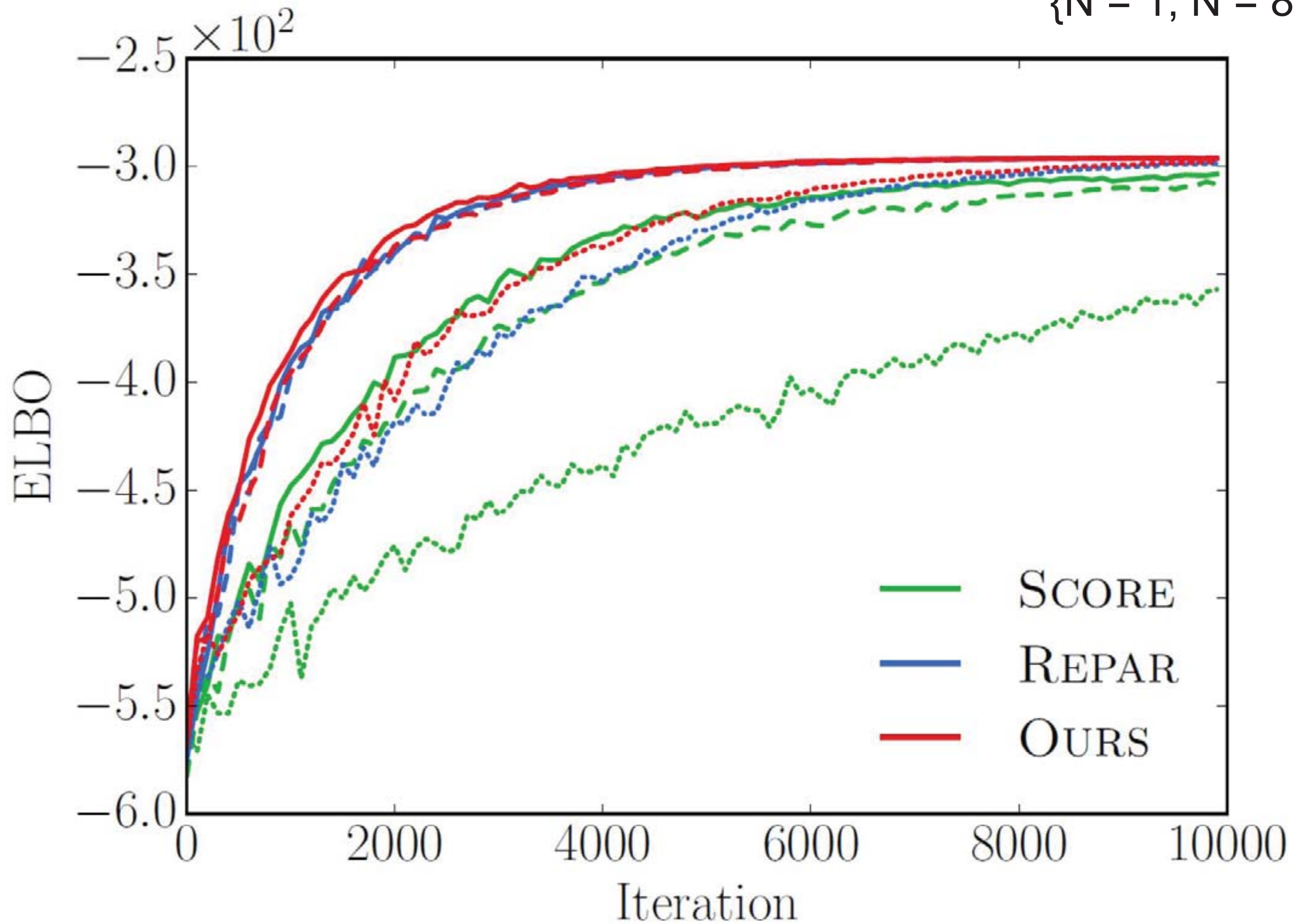
- Models random dynamics of a **controller** that tries to keep room temp. stable.

- **Non-differentiable** part: on/off of air conditioner, on which evolution of room temp. depends.

- Given noisy observations of temp. at each step, infer **on/off status** of the controller at each step.

ELBO

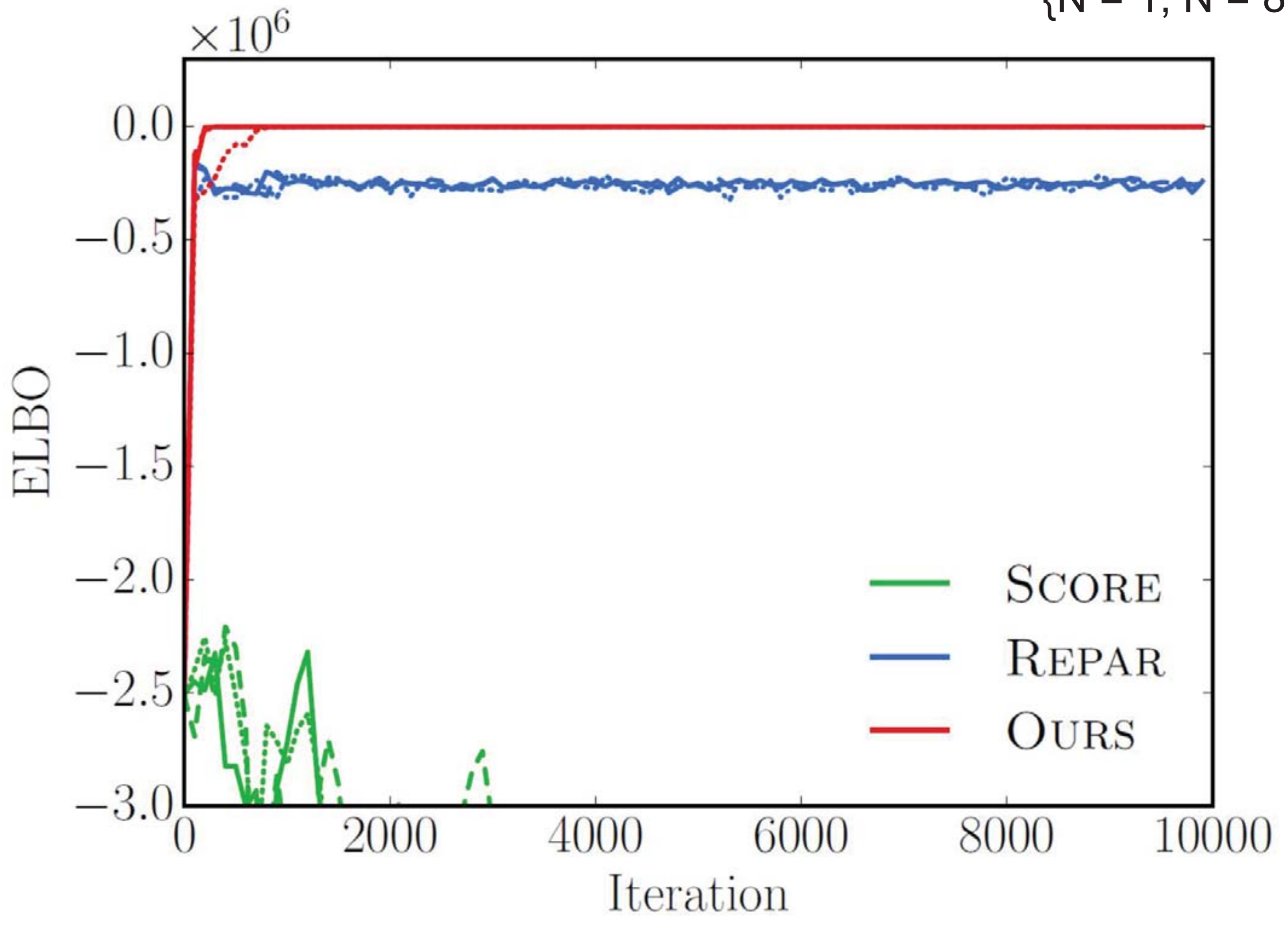
{dotted, dashed, solid} lines:
{N = 1, N = 8, N = 16}



textmsg (stepsize = 0.001)

ELBO

{dotted, dashed, solid} lines:
{N = 1, N = 8, N = 16}



temperature (stepsize = 0.01)

Computation time

Estimator	`temperature`	`textmsg`	`influenza`
SCORE	21.7	4.9	18.7
REPARAM	46.1	15.4	251.4
OURS	79.2	24.9	269.8

High-level message

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x = \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

- Careful when exchanging gradient and integration.

High-level message

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x = \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

- Careful when exchanging gradient and integration.
- May fail unexpectedly.

High-level message

$$\nabla_\theta \int H(\theta, x)\mathrm{d}x = \int \nabla_\theta H(\theta, x)\mathrm{d}x$$

- Careful when exchanging gradient and integration.
- May fail unexpectedly.
- May hold unexpectedly, but with correction.

Any questions?