Reparameterization Gradient for Non-differentiable Models from Probabilistic Programming

Hongseok Yang
KAIST, South Korea

Joint with Wonyeol Lee and Hangyeol Yu

Reparameterization Gradient for Non-differentiable Models from Probabilistic Programming

Hongseok Yang
KAIST, South Korea

Joint with Wonyeol Lee and Hangyeol Yu

Reparameterization Gradient for Non-differentiable Models from Probabilistic Programming

Hongseok Yang
KAIST, South Korea

Joint with Wonyeol Lee and Hangyeol Yu

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

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

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

High-level message

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

$$+ \text{ CorrectionTerm}$$

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

Results informally with one simple example

```
(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)
```

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

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

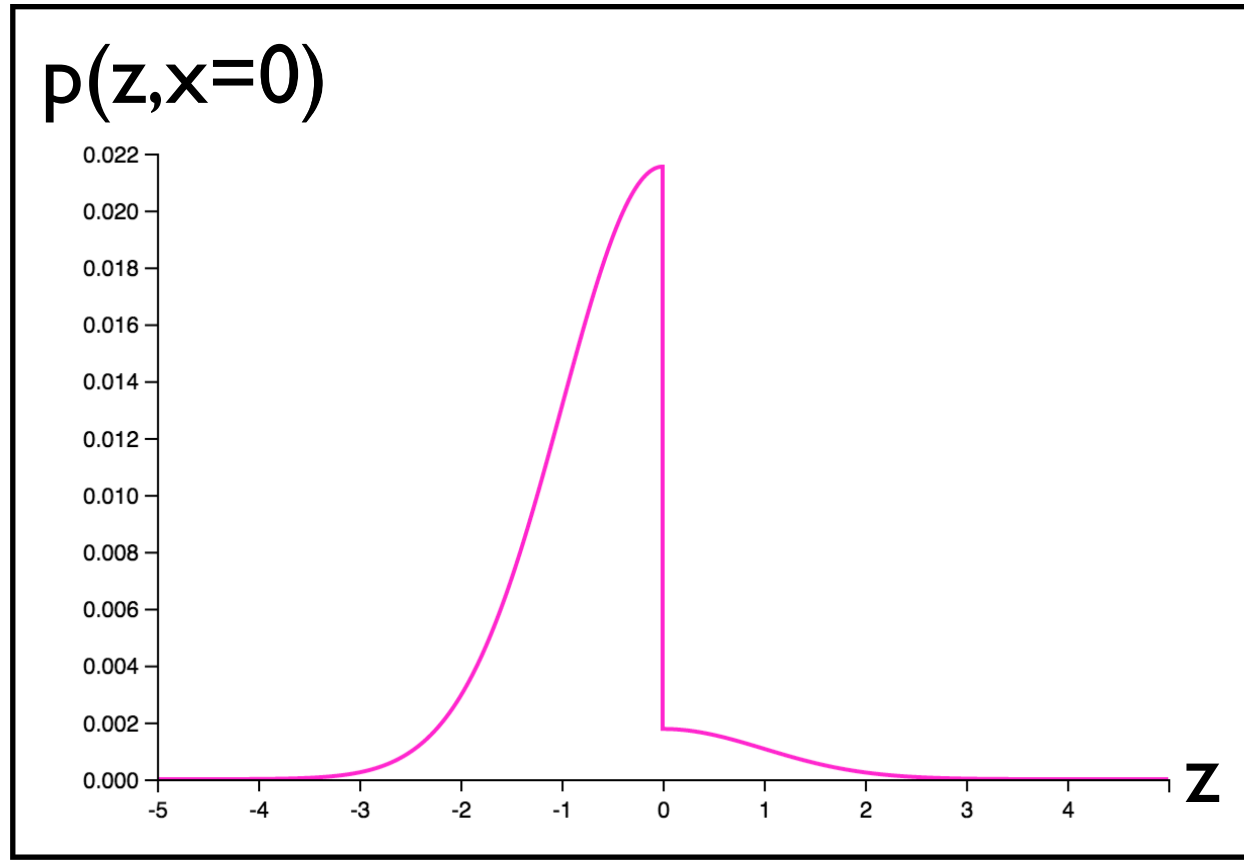
$p(z,x=0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

```
(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)
```

$\approx$

```
(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)
```

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

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

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

```
(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)
```

$\approx$

```
(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)
```

$\approx$

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

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

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

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

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

$z = \varepsilon+\theta$

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

$\approx$

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

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

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

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

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

$z = \varepsilon+\theta$

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)
```

$\approx$

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

```
(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)
```

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

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

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

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

$z = \varepsilon+\theta$

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.

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

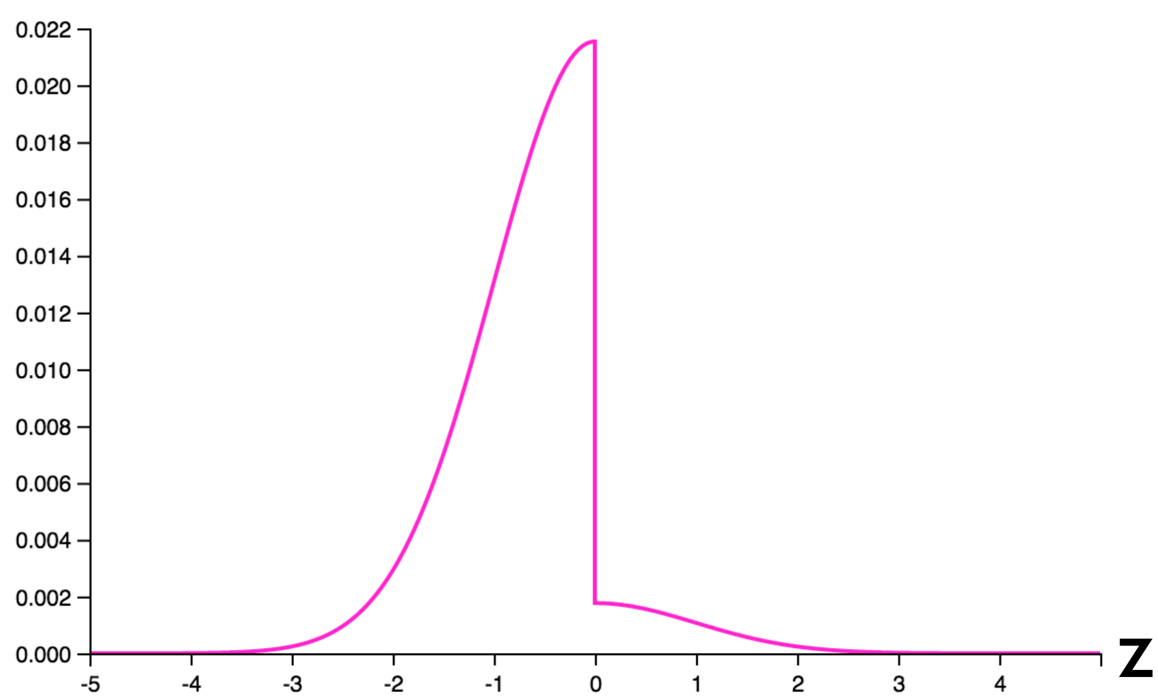
$$\boxed{\begin{aligned}&q(\varepsilon) = \mathcal{N}(\varepsilon|0,1)\\ &z = \varepsilon+\theta\end{aligned}}$$

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

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

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

p(z,x=0)



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

radient ascent on ELBO_θ.

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

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

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

$z = \varepsilon+\theta$

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

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

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

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

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

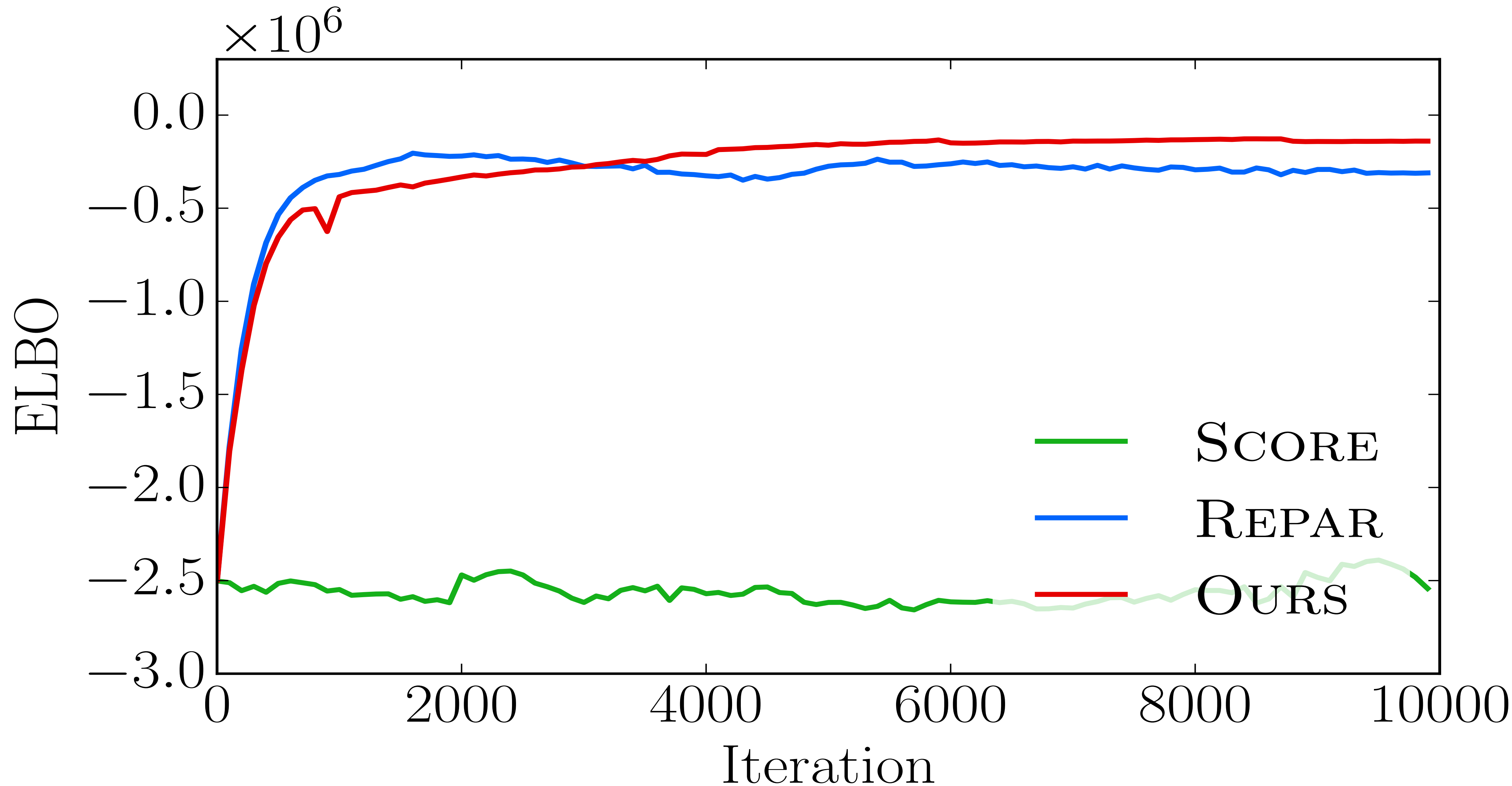
$p(z,x{=}0) = [z>0]r_1(z) + [z\leq 0]r_2(z)$

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

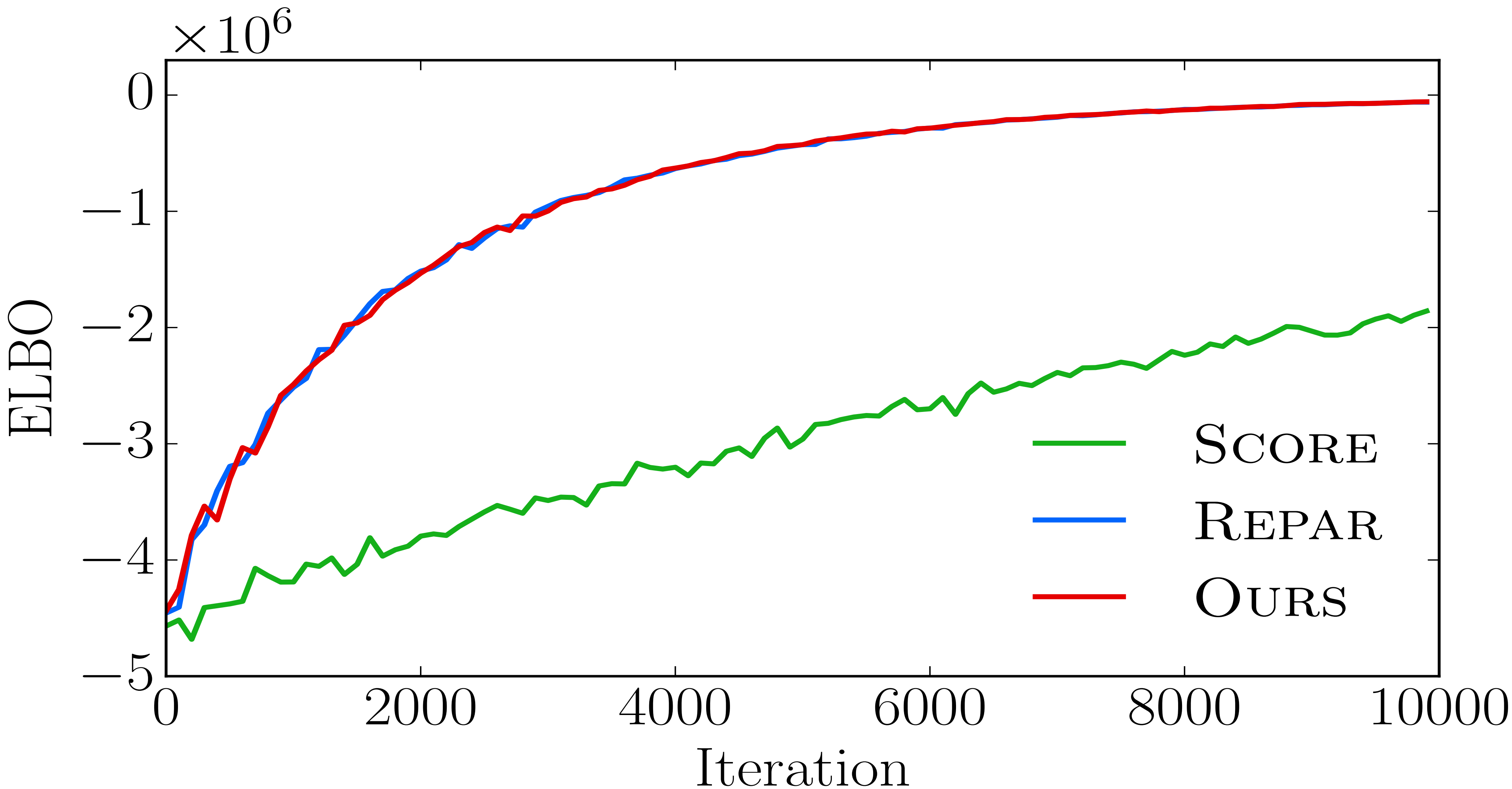
$z = \varepsilon+\theta$

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

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



(a) `temperature` (stepsize = 0.001)



(e) `influenza` (stepsize = 0.001)

Results formally

Reparameterization gradient for non-differentiable models

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

Reparameterization gradient for non-differentiable models

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

$\text{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}[\ \sum_k [\varepsilon \in f_\theta^{-1}(A_k)] \times H_k(\varepsilon,\theta)\ ]$

Reparameterization gradient for non-differentiable models

Model $\quad p(z, x^0) = \sum_k [z \in A_k] \times r_k(z).$

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

$$\nabla_\theta \mathrm{ELBO}_\theta = \mathbb{E}_{q(\varepsilon)}\Big[\sum_k [\varepsilon \in f_\theta^{-1}(A_k)] \times \nabla_\theta H_k(\varepsilon, \theta) \Big]$$

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

Accounts for the impact of moving the boundaries. Can be estimated by (optimised) manifold sampling when boundaries are affine.

$$\text{Model} \quad p(z, x^0) = \textstyle\sum_k [z \in A_k] \times r_k(z).$$

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

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

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

Correction term for k

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

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

- $\mathbf{V}(\varepsilon,\theta)_{ij} = (\partial f_\theta^{-1} / \partial \theta_i)_j$
- $\boldsymbol{\Sigma}$ is a normal vector of ∂A_k
- $\mathbf{V}(\varepsilon,\theta) \bullet \mathrm{d}\boldsymbol{\Sigma}$ is a matrix-vector product

Correction term for k

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

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

- $\mathbf{V}(\varepsilon,\theta)_{ij} = (\partial f_\theta^{-1} / \partial \theta_i)_j$
- $\boldsymbol{\Sigma}$ is a normal vector of ∂A_k
- $\mathbf{V}(\varepsilon,\theta) \bullet \mathrm{d}\boldsymbol{\Sigma}$ is a matrix-vector product

Two ingredients

- Differentiation under moving domain:

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

- Divergence theorem:

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

Reference

- “Reparameterization gradient for non-differentiable models” by Lee, Yu and Yang. 2018.