Score functions

Xun Zheng
xzheng1@andrew.cmu.edu

1 Score functions

1.1 Fisher score function

- Random variable x has density $f_\theta(x)$ parameterized by θ. Assume the support $\mathcal{X}$ is independent of θ.
- Score function measures the sensitivity of log-likelihood $\log f_\theta(x)$ to its parameter θ:

$$s(\hat{\theta}, x) \coloneqq \frac{\partial \log f_{\hat{\theta}}(x)}{\partial \theta} = \frac{1}{f_{\hat{\theta}}(x)} \frac{\partial f_{\hat{\theta}}(x)}{\partial \theta} \tag{1}$$

  Note that it is a function of both θ and x.

- Mean of score is zero:

$$\mathbb{E}_{x \sim f_{\hat{\theta}}}[s(\hat{\theta}, x)] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} f_{\hat{\theta}}(x) \, \mathrm{d}x = \frac{\partial}{\partial \theta} \underbrace{\int_{\mathcal{X}} f_{\hat{\theta}}(x) \, \mathrm{d}x}_{=1} = 0 \tag{2}$$

  Since the domain $\mathcal{X}$ is independent of θ, by Leibniz rule the integral and differentiation can be switched. Note that θ for expectation and score function has to be the same.

- Variance of score is the Fisher information:

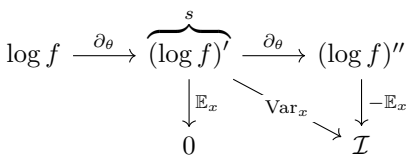
$$\mathcal{I}(\hat{\theta}) \coloneqq \mathrm{Var}_{x \sim f_{\hat{\theta}}}[s(\hat{\theta}, x)] = -\mathbb{E}_{x \sim f_{\hat{\theta}}}\big[\frac{\partial^2}{\partial \theta^2} \log f_{\hat{\theta}}(x)\big] \tag{3}$$

  which is also negative average Hessian of log density. To see why,

$$\text{LHS} = \mathbb{E}_{x \sim f_{\hat{\theta}}}[s(\hat{\theta}, x)^2] - (\underbrace{\mathbb{E}_{x \sim f_{\hat{\theta}}}[s(\hat{\theta}, x)]}_{=0})^2 \tag{4}$$

$$\text{RHS} = \mathbb{E}_{x \sim f_{\hat{\theta}}}\Big[\Big(\underbrace{\frac{1}{f_{\hat{\theta}}(x)} \frac{\partial f_{\hat{\theta}}(x)}{\partial \theta}}_{=s(\hat{\theta}, x)}\Big)^2\Big] - \underbrace{\mathbb{E}_{x \sim f_{\hat{\theta}}}\Big[\frac{\partial^2 f_{\hat{\theta}}(x)}{\partial \theta^2} \frac{1}{f_{\hat{\theta}}(x)}\Big]}_{=0 \text{ by Leibniz}} \tag{5}$$

- Informal summary:



1.2 Hyvärinen score function

- Hyvärinen score measures the sensitivity of log-likelihood w.r.t. x:

$$h_{f_\theta}(\hat{x}) \coloneqq \frac{\partial \log f_\theta(\hat{x})}{\partial x} = \frac{1}{f_\theta(\hat{x})} \frac{\partial f_\theta(\hat{x})}{\partial x} \tag{6}$$

- Hyvärinen (2005) phrased this as the Fisher score w.r.t. a hypothetical location parameter: define a new density $f_{\theta,\mu}(x) \coloneqq f_\theta(x-\mu)$, and let $\hat{\mu}=0$, then the Fisher score of $f_{\theta,\mu}$ restricted to μ is

$$s(\hat{\mu},\hat{x}) = \frac{\partial \log f_\theta(\hat{x}-\hat{\mu})}{\partial \mu} = \frac{1}{f_\theta(\hat{x}-\hat{\mu})} \frac{\partial f_\theta(\hat{x}-\hat{\mu})}{\partial (x-\mu)} \frac{\partial(\hat{x}-\hat{\mu})}{\partial \mu} = -h_{f_\theta}(\hat{x}) \tag{7}$$

  Caution: for some distributions μ may affect the support $\mathcal{X}$, such as exponential distribution, in which case the regularity condition for (2) does not hold anymore:

$$f_\theta(x) = \theta e^{-\theta x} \cdot \mathbb{I}(x>0) \implies \mathbb{E}_{x\sim f_\theta}[h_{f_\theta}(x)] = -\theta \neq 0 \tag{8}$$

- Hyvärinen score is agnostic about the partition function. Suppose

$$f_\theta(x) = \frac{1}{Z_\theta}\tilde{f}_\theta(x), \quad Z_\theta = \int_{\mathcal{X}} \tilde{f}_\theta(x)\,\mathrm{d}x \tag{9}$$

  then the Hyvärinen score is just the sensitivity of the log-unnormalized-density w.r.t. x:

$$h_{f_\theta}(\hat{x}) = \frac{\partial\big[\log \tilde{f}_\theta(\hat{x}) - \log Z_\theta\big]}{\partial x} = \frac{\partial \log \tilde{f}_\theta(\hat{x})}{\partial x} \tag{10}$$

- Score matching can be useful in the following parametric density estimate problem: given i.i.d. samples from some unknown density f^*, estimate a parametric density f_θ, only using evaluations of $\tilde{f}_\theta$.

- Score matching density estimator: $\hat{\theta} = \operatorname{argmin}_\theta J(\theta)$, where

$$J(\theta) = \mathbb{E}_{x\sim f^*}\big[\frac{1}{2}\left\|h_{f_\theta}(x) - h_{f^*}(x)\right\|_2^2\big] \tag{11}$$

$$= \mathbb{E}_{x\sim f^*}\big[\frac{1}{2}\left\|h_{f_\theta}(x)\right\|_2^2 + \underbrace{\frac{1}{2}\left\|h_{f^*}(x)\right\|_2^2}_{\text{independent of }\theta} - \underbrace{\langle h_{f_\theta}(x), h_{f^*}(x)\rangle}_{\alpha}\big] \tag{12}$$

$$= \mathbb{E}_{x\sim f^*}\big[\operatorname{tr}(H) + \frac{1}{2}\left\|\mathbf{h}\right\|^2\big] + \text{constant} \tag{13}$$

  with shortcuts

$$\mathbf{h} = h_{f_\theta}(x), \quad H = \frac{\partial h_{f_\theta}(x)}{\partial x} = \frac{\partial^2 \log f_\theta(x)}{\partial x \partial x} \tag{14}$$

$$\log f \xrightarrow{\partial_x} \underbrace{(\log f)'}_{\mathbf{h}} \xrightarrow{\partial_x} \underbrace{(\log f)''}_{H}$$

  To see how f^* disappears, (roughly)

$$\mathbb{E}[\alpha] = \int f^*(x) \frac{\partial \log f_\theta(x)}{\partial x} \frac{\partial \log f^*(x)}{\partial x}\,\mathrm{d}x \tag{15}$$

$$= \int f^*(x) \frac{\partial \log f_\theta(x)}{\partial x} \big(\frac{1}{f^*(x)} \frac{\partial f^*(x)}{\partial x}\big)\,\mathrm{d}x \tag{16}$$

$$= \int \frac{\partial \log f_\theta(x)}{\partial x} \frac{\partial f^*(x)}{\partial x}\,\mathrm{d}x \tag{17}$$

$$= -\int \frac{\partial^2 \log f_\theta(x)}{\partial x \partial x} f^*(x)\,\mathrm{d}x \tag{18}$$

Last line follows from integration by parts and diminishing tails: $\int gf' = [gf]_{-\infty}^{+\infty} - \int g'f$.

Under regularity conditions, $J(\theta) = 0 \iff \theta = \theta^*$. Since $J(\theta) \geq 0$, this motivates the score matching density estimator.

- Pros: $J(\theta)$ can be optimized using only gradient and Hessian of $\log \tilde{f}_\theta(x)$, no estimates of Z_θ and f^* are involved.
- Cons: need to know how θ behaves in $\mathbf{h}$ and $\mathrm{tr}(H)$, *e.g.*, access to $\frac{\partial \,\mathrm{tr}(H)}{\partial \theta}$, $\frac{\partial \mathbf{h}}{\partial \theta}$.

References

A. Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 2005.