Score Matching, SURE, and Tweedie's formula

Tzu-Mao Li

December 2024

1 Score matching

Given data $x_0, x_1, \cdots \in \mathbb{R}^n$, such that $x \sim p(x)$ where p is some unknown probability density function, we want to fit their *score function*:

$$\Psi(x) = \nabla_x \log p(x), \tag{1}$$

where ∇_x denotes the gradient w.r.t. x. One reason we might be interested in this is that we might want to apply (Metropolis-adjusted) Langevin Monte Carlo or even Hamiltonian Monte Carlo to sample from the probability density p (e.g., [SE19]).

To find Ψ, we use a parametric function $\hat{\Psi}(x;\theta)$ parameterized by θ and find θ to minimize the mean square error difference [HD05]. However, directly minimizing the mean square error is tricky, since Ψ is unknown – we are only given a bunch of data x_i! However, turns out we can rewrite the mean square error to a form that is completely avoid of the target score function Ψ:

$$\begin{aligned}
&\arg\min_\theta E_{x\sim p(x)}\left[\left\|\hat{\Psi}(x;\theta) - \Psi(x)\right\|^2\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[-2\hat{\Psi}(x;\theta)\cdot\frac{\nabla_x p(x)}{p(x)} + \left\|\hat{\Psi}(x;\theta)\right\|^2\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[2\nabla_x\cdot\hat{\Psi}(x;\theta) + \left\|\hat{\Psi}(x;\theta)\right\|^2\right],
\end{aligned} \tag{2}$$

where $\nabla_x\cdot$ denotes the divergence w.r.t. x.

The last equation is true because of integration by parts and divergence theorem:

$$\int_{\mathbb{R}^n}\hat{\Psi}(x;\theta)\cdot\nabla_x p(x)\mathrm{d}x = -\int_{\mathbb{R}^n}\nabla_x\cdot\hat{\Psi}(x;\theta)p(x)\mathrm{d}x. \tag{3}$$

(assuming p diminishes to zero at infinity.)

Thus we can find a parametric function $\hat{\Psi}$ that minimizes the last loss function to optimally approximate the target score function.

2 Stein's Unbiased Risk Estimate (SURE)

Turns out the derivation above is closely related to SURE [Ste81] and Tweedie's formula [Rob92]. If we assume $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x-\mu\|^2}{2}}$ (i.e., a unit variance normal distribution), and we want to find an *optimal denoiser* $f(x;\theta)$ to find the mean μ given the same data, turns out we can also find the optimal denoiser in the mean square error sense without knowning the target μ:

$$\begin{aligned}
&\arg\min_\theta E_{x\sim p(x)}\left[\|f(x;\theta)-\mu\|^2\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[\|f(x;\theta)\|^2 - 2\mu\cdot f(x;\theta)\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[\|f(x;\theta)-x\|^2 - 2(\mu-x)\cdot f(x;\theta)\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[\|f(x;\theta)-x\|^2 + 2\nabla_x\cdot f(x;\theta)\right] \\
=&\arg\min_\theta E_{x\sim p(x)}\left[\|g(x;\theta)\|^2 + 2\nabla_x\cdot g(x;\theta)\right]
\end{aligned} \tag{4}$$

where $g(x;\theta) = f(x;\theta) - x$. Intuitively, g is the estimation of the noise, assuming x is obtained by adding some noise to the latent image.

The second last equation is true, again, because of integration by parts and divergence theorem:

$$\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x-\mu\|^2}{2}} (\mu - x)\cdot f(x;\theta)\mathrm{d}x = -\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x-\mu\|^2}{2}} \nabla_x \cdot f(x;\theta)\mathrm{d}x. \tag{5}$$

3 Optimal noise estimation is the optimal score function

Now, if we simply do a pattern matching by comparing the last equalities in Equation (2) and (4), the optimal noise estimate g (more precisely, the "offset" of the optimal denoiser $f - x$) is the same as the optimal score function $\hat{\Psi}$!

What does this mean? This means that one (deterministic) step of Langevin Monte Carlo $x + \hat{\Psi}(x)$ is *exactly* equivalent to optimally denoising x by pretending x is corrupted by a unit variance Gaussian (aka Tweedie's formula). For Gaussians with an arbitrary covariance matrix, the equivalence can be shown by linearly transforming the score function $\hat{\Psi}$ using the covariance matrix. So any parametric function f or $\hat{\Psi}$ that is very good at denoising, will be great at moving your data closer to their typical modes of the underlying data distribution.

The same observation has already been made by earlier works [Men+21; Luo22], so this is nothing new. But hopefully this document is less dense!

References

[HD05] Aapo Hyvärinen and Peter Dayan. "Estimation of non-normalized statistical models by score matching." In: *Journal of Machine Learning Research* 6.4 (2005).

[Luo22] Calvin Luo. "Understanding diffusion models: A unified perspective". In: *arXiv preprint arXiv:2208.11970* (2022).

[Men+21] Chenlin Meng et al. "Estimating high order gradients of the data distribution by denoising". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25359–25369.

[Rob92] Herbert E Robbins. "An empirical Bayes approach to statistics". In: *Breakthroughs in Statistics: Foundations and basic theory*. Springer, 1992, pp. 388–394.

[SE19] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in neural information processing systems* 32 (2019).

[Ste81] Charles M Stein. "Estimation of the mean of a multivariate normal distribution". In: *The annals of Statistics* (1981), pp. 1135–1151.