Thoughts On: Tweedie's Formula

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

Statistical inference fundamentally concerns how to estimate unknown parameters with observable data. Tweedie's formula, an identity, solves this challenge elegantly by establishing an exact relationship $\sigma^2\times$ the derivative of the marginal log-likelihood. This fundamental formula has profound implications and insights for both theoretical statistics and machine learning.

1 Summary

Tweedie's formula is a bridge between statistical inference and machine learning. It states that the posterior expectation must equal the observation plus $\sigma^2 \times \frac{d}{dx}\log(m(x))$, whee σ^2 is the variance of the Gaussian noise. This formula effectively determines how much to trust individual observations when compared with the general population distribution. Its most far reaching implication has been its recent rediscovery in denoising autoencoders and diffusion models, where it formed the underlying backbone of the architecture, leading to the optimal Bayesian solution. This conclusion has recently gained wide spread attention in various generative AI communities such as diffusion modeling. As a result I have decided to write some of my thoughts and insights on this formula to hopefully provide a more holistic understanding in a concise format. These recent connections reveal novel insights on an old discovery regarding the fundamental principles of statistics and machine learning opening the opportunities for better models.

2 Tweedie's Formula

Tweedie's formula describes that under appropriate regularity conditions, and if we have a parameter θ with some prior distribution $g(\theta)$ and an observation X which follows the distribution $f(x|\theta)$, then:

$$\mathbb{E}[\theta|X=x] = x + \sigma^2 \frac{d}{dx} \log m(x)$$

Here, $X = \theta + \epsilon, \epsilon = N(0, \sigma^2)$ and, $m(x) = \int f(x|\theta)g(\theta)d\theta$ is the marginal density of the variable X. This simple equation essentially states that the posterior expectation equals the observation plus an adjustment term. This adjustment is determined by the derivative of the log marginal density. Which fundamentally describes how much one needs to reduce the estimates based on the general population distribution.

3 Theoretical Foundations

3.1 The Bayesian Perspective

From a classic Bayesian perspective, Tweedie's formula provides a clean computational shortcut. Rather than applying Bayes' theorem to obtain the posterior distribution:

$$p(\theta|x) = \frac{f(x|\theta)g(\theta)}{m(x)}$$

and then integrating to find the posterior mean:

$$\mathbb{E}[\theta|X = x] = \int \theta p(\theta|x) d\theta$$

The shortcut holds exactly when the observation model is additive Gaussian with known variance under σ^2. Tweedie's formula gives us a direct path to the posterior expectation through differentiation rather than integration cutting the computation down substantially. This is particularly useful since integration is more expensive than differentiation and since it allows for the posterior distribution to be complex while $m(x)$ is tractable.

3.2 Exponential Families

For exponential family distributions, Tweedie's formula takes a particularly elegant form as demonstrated below:

$$f(x|\theta) = h(x) \exp(\theta T(x) - A(\theta))$$

If the equation above is true, then we can also connect Tweedie's formula to convex analysis:

$$\mathbb{E}[\theta|X = x] = \nabla_\psi b^*(\psi) = \hat{\theta}_{MLE}(\psi)$$

This gradient returns the natural (or MLE/MAP) parameter; it coincides with the posterior mean only under a flat prior and large-sample limits. Here $\psi = T(x)$ and b^* is the convex conjugate of the log-partition function $A(\theta)$. This convex conjugate is defined as:

$$b^*(\psi) = \sup_\theta \{\theta^T \psi - A(\theta)\}$$

This is a deep connection between Bayesian inference and convex optimization. The gradient of the convex conjugate $\nabla_\psi b^*(\psi)$ corresponds exactly to the parameter θ that maximizes $\theta^T \psi - A(\theta)$, which is precisely the posterior expectation we desire. This mathematical relationship demonstrates how some divergences arise in Bayesian inferencing which also allows for geometrical interpretations of statistical estimation methods.

4 Deriving Statistical Estimators

To demonstrate the universality of Tweedie's formula, lets derive an estimator in statistics.

Consider the example where we define the following variables $X|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(0, \tau^2)$ and the marginal distribution to be $X \sim N(0, \sigma^2 + \tau^2)$. Here τ defines the weight of prior knowledge when compared to observed data, or uncertainty.

We apply Tweedie's formula in the following form:

1. Calculate $m(x) = (2\pi(\sigma^2 + \tau^2))^{-1/2} \exp(-x^2/(2(\sigma^2 + \tau^2)))$
2. Find $\log m(x) = -1/2 \log(2\pi(\sigma^2 + \tau^2)) - x^2/(2(\sigma^2 + \tau^2))$
3. Compute the derivative: $d/dx \log m(x) = -x/(\sigma^2 + \tau^2)$
4. Apply Tweedie's formula: $E[\theta|X = x] = x - x/(\sigma^2 + \tau^2) = (\tau^2/(\sigma^2 + \tau^2))x$

This recovers the James-Stein estimator, which shrinks the maximum likelihood estimate x toward zero by a factor of $\tau^2/(\sigma^2 + \tau^2)$. This elegance demonstrates how Tweedie's formula naturally produces optimal shrinkage estimators.

5 Machine Learning

5.1 Denoising Autoencoders

Tweedie's formula is fundamental to denoising autoencoders. When we add Gaussian noise $\varepsilon \sim N(0, \sigma^2 I)$ to some input data x, and train a neural network to reconstruct the original input, the optimal reconstruction function becomes:

$$r^*(x + \varepsilon) = (x + \varepsilon) + \sigma^2 \nabla \log p(x + \varepsilon)$$

This is isomorphic to Tweedie's formula, revealing that denoising autoencoders implicitly perform Bayesian inference at their core. This illustrates why these types of architectures are so effective at generating representations which estimate the gradient of log density, or score functions, of the data distribution.

5.2 Score-Based Generative Models

Similarly, denoising autoencoders, generative modeling, and particularly diffusion models, leverage score functions innately. In diffusion models in particular, Tweedie's formula arises because diffusion aims to reverse flow of information gathered through the process of Bayesian inference. For example, here is the score function for diffusion models, defined as the gradient of the log probability density:

$$\nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$$

This gradient guides the generative process, allowing models to transform noise into structured data. Tweedie's formula is a fundamental property of denoising diffusion models because these process flows are effectively reversing the process of Bayesian inference.

- **Forward process:** $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$ and α_t decreases with time
- **Reverse process:** $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t))$ where ϵ_θ is a learned noise predictor

The optimal noise predictor $\epsilon_\theta(x_t, t)$ seeks to estimate $\mathbb{E}[\epsilon | x_t]$. Through a slight change in the variables, we can equate this to estimating $\nabla_{x_t} \log p(x_t)$. Which is, in other words, the precise the score function.

We can then describe the reverse process in terms of probabilities:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$$

where,

$$\mu_\theta(x_t.t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t))}$$

With $s_\theta, \sim \nabla_{x_t} \log p_{\log(x_t)}$ resulting in a complete recovery of Tweedie's adjustment $\sigma_t^2 s_\theta(x_t, t)$The Bayesian inference is thus a fundamental part of diffusion models.

5.3 Posterior Variance

Tweedie's formula can also be extended to higher moments, demonstrating direction relationships to information theory. The posterior variance is given by:

$$\textbf{Var}(\theta | X = x) \sigma^2 + \sigma^4 \frac{d^2}{dx^2} \log(m(x))$$

This expression directly relates to Fisher information, denoted $I(x)$, which quantifies how much information an observation x carries about the unknown parameter θ. Specifically:

$$\text{Var}(\theta|X = x) = I(x)^{-1}$$

In this way, when we consider the second derivative, we can see that regions of descending steepness in the log density will provide higher information certainty, while flatter regions will provide less certainty. Think of Fisher information as the ”steepness” of the likelihood landscape, and that the steeper the curve, the more precise the estimate. This works in high dimensional spaces while providing confidence in accuracy while significantly reducing the computational costs.

5.4 Multivariate

Similarly, in the multivariate settings, Tweedie's formula can be generalized to:

$$E[\theta|X = x] = x + \Sigma\nabla_x \log m(x)$$

where $\nabla_{\mathbf{x}}$ is the gradient respective of $\mathbf{x}$ and Σ is the noise covariance.

Here the vectors and observations are handled in a manner which has various applications in high-dimensional problem spaces. In each case, the multivariate formulation can provide a direct way to reduce higher dimensions, which is determined by the gradient of the log marginal density.

6 Conclusion

Tweedie's formula is a very fundamental equation in statistical inferencing that demonstrates relationships in various other fields and applications. These include empirical Bayes, information geometry, and modern machine learning where Tweedie's formula offers both a theoretically sound foundation and offers practical guidance on estimation problems. The local behavior of the marginal distribution contains precisely the granular information necessary for optimal Bayesian estimation, resulting in a a universality between Tweedie's formula reducing statistical estimators and denoising in modern machine learning.