Andrej Leban
University of Michigan

Energy-Tweedie: Score meets Score, Energy meets Energy

A general form of Tweedie's formula

PAPER CODE

1 BACKGROUND

Denoising & Tweedie's formula (1956)

Data $X \sim p$ noised with $\epsilon \sim q$: $Y = X + \epsilon$. $m = p * q$ – the marginal, and $s_m = \nabla_y \log m$ – the *score*. Under *Gaussian* noise ($q = \mathcal{N}(0, \Sigma)$) the **mean** of the *denoising posterior* $P(X \mid Y = y)$ gives the score:

$$s_m(y) \;=\; \Sigma^{-1}(\mathbb{E}_X[X \mid Y = y] - y)$$

This is the identity behind diffusion models, and the posterior mean can be learned with the *squared loss* (MSE). **What if the noise is not Gaussian?**

Gibbs noise

$$q_\eta(u) \;\propto\; e^{-c\,\mathcal{E}(u)},$$

with *noise potential* $\mathcal{E}$, *scale* c, and parameters η. Also called an *energy-based model*. Gaussian, Laplace, generalized Gaussian, Student-t, … all fit this form.

Scoring rules

Scoring rules have traditionally been used to evaluate *forecasts*: they compare the *predictive distribution* P of a forecaster against a sample y from the ground truth. *Proper* scoring rules are minimized by the ground truth distribution, so they also work as *distributional training losses*.

For some *discrepancy* ρ, a general **kernel score** is:

$$S_\rho(P, y) \;=\; \mathbb{E}_{X\sim P}[\rho(X, y)] - \tfrac{1}{2}\mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P}[\rho(X, X')]$$

An example is the *energy score*: $\mathrm{ES}_{l_d,\beta}(P, y) = \mathbb{E}_{X\sim P}[\|X - y\|^\beta] - \frac{1}{2}\mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P}[\|X - X'\|^\beta]$, i.e., $\rho(x, y) = \|x - y\|^\beta$.

2 IDENTITY

THE ENERGY-TWEEDIE IDENTITY

$$s_m(y) \;=\; -c\,\nabla_y^{\mathrm{PD}}\, S_\rho\big(P(X \mid Y = y),\, y\big),$$

where $\rho(x, y) = \mathcal{E}(y - x)$ is the discrepancy *induced* by the *noise potential* $\mathcal{E}$, and ∇_y^{PD} is the *path derivative* (*stopgrad*) with respect to y.

The **score** of the noisy marginal is thus given by the path derivative of the **kernel score** induced by the noise, at the denoising posterior.

The correspondence

The familiar correspondence between Gaussian noise, posterior means, squared loss, and Tweedie's formula is lifted to a distributional correspondence between Gibbs noise distributions, full posterior distributions, kernel scoring rules, and the Energy-Tweedie identity. As an example, we contrast plugging in the generalized Gaussian with the Gaussian:

	NOISE	OBJECTIVE	POSTERIOR OBJECT	SCORE
Gen. Gaussian	$\propto \exp\!\left(-\frac{\lambda}{\beta}\|u\|^\beta_{\Sigma^{-1}}\right)$	$\mathrm{ES}_{\Sigma^{-1},\beta}(P(X \mid Y = y), x)$	$P(X \mid Y = y)$	$-\frac{\lambda}{\beta}\nabla_y^{\mathrm{PD}}\mathrm{ES}_{\Sigma^{-1},\beta}(P(X \mid Y = y), y)$
Gaussian	$\propto \exp\!\left(-\frac{1}{2}u^\top\Sigma^{-1}u\right)$	$\|\mathbb{E}_X[X \mid Y = y] - x\|^2_{\Sigma^{-1}}$	$\mathbb{E}_X[X \mid Y = y]$	$\Sigma^{-1}(\mathbb{E}_X[X \mid Y = y] - y)$

The Generalized reduces to the Gaussian when $\beta = 2, \lambda = 1$, and the Energy-Tweedie identity recovers the classical Tweedie formula.

3 USES

Score estimation

Posterior-based:

$$\hat{s}_m(y) \;=\; -\frac{c}{N}\sum_{i=1}^{N}\nabla_y^{\mathrm{PD}}\, S_\rho(\delta_{X^{(i)}}, y), \qquad X^{(i)} \overset{\text{iid}}{\sim} P_\theta(X \mid Y = y)$$

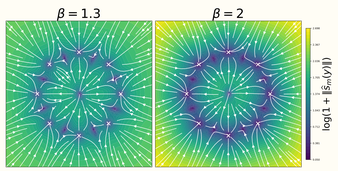
ET-DSM – Energy-Tweedie Denoising Score Matching:

$$\min_\theta\; \mathbb{E}_{X\sim p,\, Y\sim q_\eta(\cdot|X)}\left[\left\|s_\theta(Y, \eta) + c\,\nabla_Y^{\mathrm{PD}}\, S_\rho(\delta_X, Y)\right\|_2^2\right]$$

Examples from the *Eight Gaussians* dataset, noised with generalized Gaussian noise (over a range of parameters):

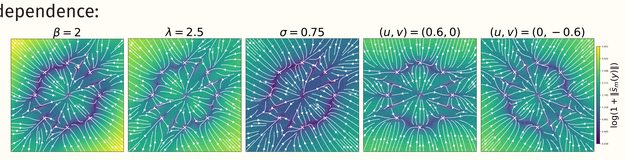
SCORE ESTIMATOR	NORM. MSE ↓	COSINE SIM. ↑
ET-Posterior	0.029–0.077	0.951–0.982
ET-DSM	0.014–0.019	0.990–0.992
Hyvärinen SM	0.022–0.027	0.987–0.989

Modifying the β parameter shows mean- vs. median-seeking behavior of the score:



Fixed $\lambda = 1.8$, $(u, v) = (0, 0)$, $\sigma = 0.42$ (left) and 0.57 (right), from the parameterization →

Varying the other parameters reveals the tail, anisotropy, and noise magnitude dependence:



Base is $(\beta, \lambda, \sigma, u, v) = (1.4, 1.8, 0.45, 0, 0.6)$, each panel changes one parameter.

Noise parameter estimation

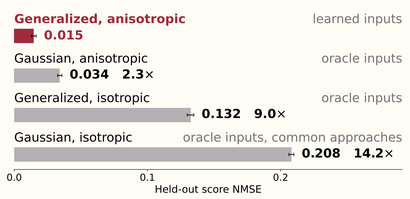
Use *another score estimator* $\hat{s}_m$ and turn it into an *M-estimation* problem:

$$\min_\eta\; \mathbb{E}_{Y\sim m}\left[\left\|\hat{s}_m(Y) + c\,\nabla_Y^{\mathrm{PD}} S_\rho(P_\theta(X \mid Y), Y)\right\|_2^2\right]$$

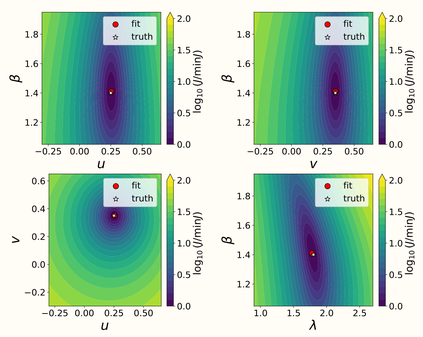
Using generalized Gaussian noise with:

$$\eta = (\beta, \lambda, u, v), \qquad \Sigma \;=\; \sigma^2\begin{pmatrix} 1+u & v \\ v & 1-u \end{pmatrix},$$

we compare the full ET approach to those with restricted distributions *that we give oracle inputs*:



The parameter values are recovered almost exactly in a single fit:



Energy-Tweedie diffusion

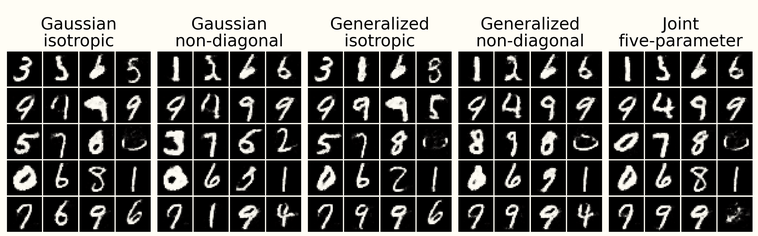
Energy-Tweedie diffusion enables *paths* through the *parameters of the noise distribution* $t \mapsto \eta_t$ instead of a scalar *diffusion schedule*. This leads to a **path-induced training loss:**

$$\min_\theta\; \mathbb{E}_{t\sim\mathrm{Unif}([0,1]),\, X\sim p,\, Y\sim q_{\eta_t}(\cdot|X)}\Big[S_{\rho_t}(P_\theta(X \mid Y, t),\; X)\Big],$$

and/or to **sampling from a freely chosen path**:

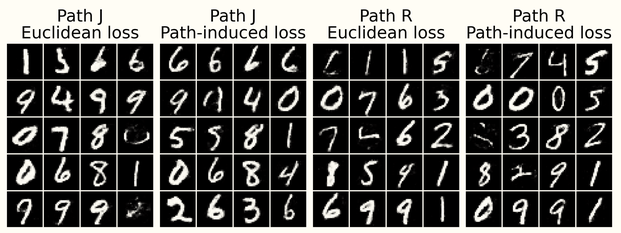
Gaussian: $(\beta, \lambda) = (2, 1)$ generalized: $(\beta, \lambda) = (1.4, 1.8)$ isotropic: $(u, v) = (0, 0)$ non-diagonal: $(u, v) = (0, 0.6)$
Joint five-parameter: $\beta_t\colon 2 \to 1.4$, $\lambda_t\colon 1 \to 1.8$, $u_t\colon 0 \to 0.3 \to 0$, $v_t\colon 0 \to 0.6$; **R**eversed joint: J with $\beta_t\colon 1.4 \to 2$

Fixed training loss & model, varying the sampling path:



SAMPLING PATH	ED ↓	TV ↓
Generalized, non-diagonal	0.218	0.085
Gaussian, non-diagonal	0.247	0.111
Joint five-parameter (J)	0.303	0.107
Gaussian, isotropic	0.574	0.162
Generalized, isotropic	0.637	0.113
Real images	0.015	0.074

For a chosen path, train with the induced loss:



PATH	TRAINING LOSS	ED ↓	TV ↓
J	Path-induced	0.169	0.079
J	Euclidean	0.303	0.107
R	Euclidean	0.329	0.118
R	Path-induced	0.382	0.106

leban@umich.edu · arXiv:2512.23818 · https://github.com/andleb/et