Generative models and score matching

Samy Tindel

Purdue University

Stochastic calculus and Machine Learning - MA 595

PURDUE UNIVERSITY

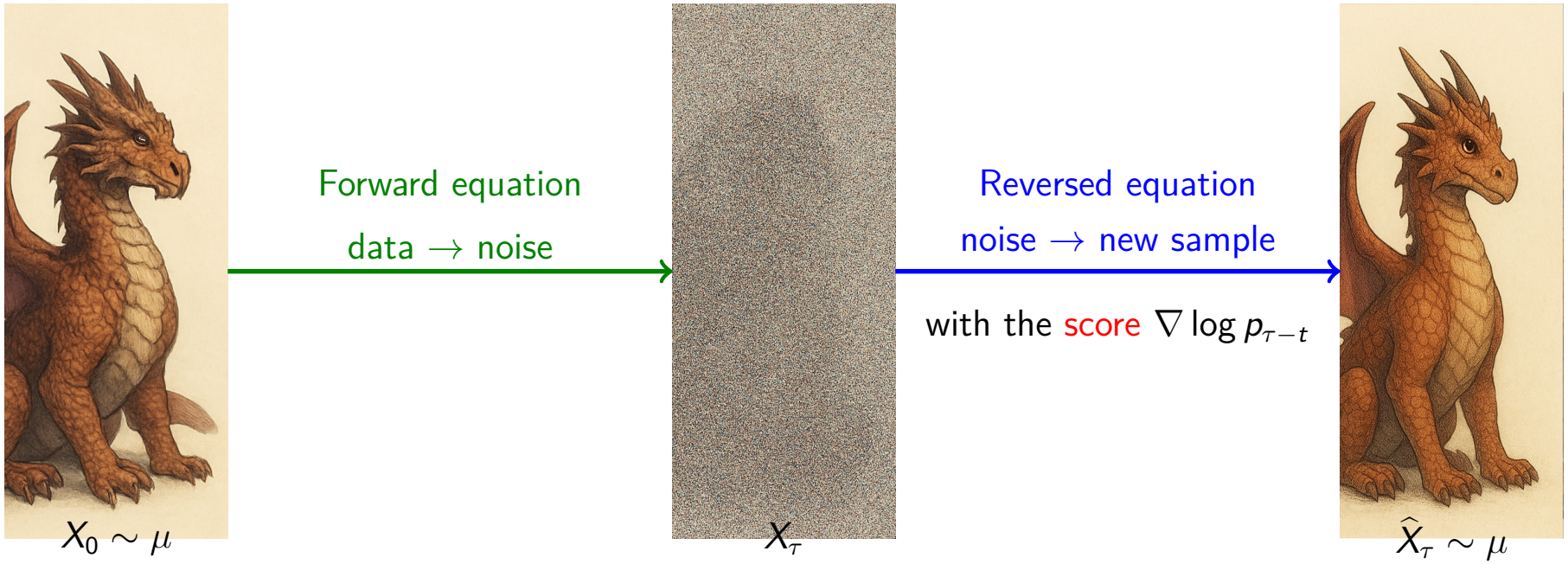
Outline

1. The generative algorithm
2. The affine forward equation
3. The score-matching identity
4. Denoising and training

Outline

1. The generative algorithm
2. The affine forward equation
3. The score-matching identity
4. Denoising and training

From noise to data



Missing ingredient: The score $\nabla \log p_{\tau-t}$ is unknown

$\hookrightarrow$ To be learned from the data: this chapter

Recall: the reversed equation

Forward equation: $X_0 \sim \mu$ arbitrary law on $\mathbb{R}^n$

$$dX_t = b(t, X_t)\, dt + \sigma(t)\, dW_t, \quad t \in [0, \tau]$$

Reversed equation: $\widehat{X}_t = X_{\tau - t}$ solves, for $t \in [0, \tau - \varepsilon]$,

$$d\widehat{X}_t = \Big[-b(\tau - t, \widehat{X}_t) + a(\tau - t)\, \nabla \log p_{\tau - t}(\widehat{X}_t)\Big]\, dt + \sigma(\tau - t)\, d\widehat{W}_t$$

Notation: $a = \sigma\sigma^*$, p_t density of X_t, $\varepsilon > 0$ cutoff

The algorithm in four steps

Data: Sample $x_1, \dots, x_N$ of an unknown law μ on $\mathbb{R}^n$

Training phase (once):

i. Noising: forward equation from each $x_i \to$ samples of p_t
ii. Training: fit a network $s_\theta(t, x) \approx \nabla \log p_t(x)$

Sampling phase (for each new sample):

iii. Initialization: $\widehat{X}_0$ from a simple law $\approx$ law of X_τ
iv. Sampling: discretized reversed equation with s_θ, stopped at $\tau - \varepsilon$

Goals of Chapters 6 and 7

Five goals:

i. Training well posed: a loss computable from the data
ii. Training tractable: regression on a Gaussian target
iii. Imperfect score: stability of the reversed equation
iv. Initialization: law of X_τ close to a fixed law
v. Discretization: rate for the Euler–Maruyama scheme

Split: (i)–(ii) this chapter, (iii)–(v) Chapter 7

First: Choose b, σ so that everything is explicit

Outline

1. The generative algorithm
2. The affine forward equation
3. The score-matching identity
4. Denoising and training

Affine setting

Assumption:

1. $b(t, x) = B(t)\, x$, with $B : [0, \tau] \to \mathbb{R}^{n \times n}$ infinitely differentiable
2. $\sigma : [0, \tau] \to \mathbb{R}^{n \times n}$ infinitely differentiable, $\sigma(t)$ invertible

Forward equation: $X_0 \sim \mu$, $X_0 \in \mathcal{F}_0$

$$dX_t = B(t)\, X_t\, dt + \sigma(t)\, dW_t$$

Examples: Ornstein–Uhlenbeck (B, σ constant)
$\hookrightarrow$ Variance preserving: $B = -\frac{1}{2}\beta \,\mathrm{Id}$, $\sigma = \sqrt{\beta}\,\mathrm{Id}$

Resolvent

Definition 1

Let B as in the affine assumption. The resolvent $\mathcal{R} : [0, \tau] \to \mathbb{R}^{n \times n}$ is the solution of

$$\dot{\mathcal{R}}(t) = B(t)\,\mathcal{R}(t), \quad \mathcal{R}(0) = \mathrm{Id}$$

Properties: Linear ODE theory

- $\mathcal{R}$ infinitely differentiable, $\mathcal{R}(t)$ invertible for all t
- $\frac{d}{dt}\,\mathcal{R}(t)^{-1} = -\mathcal{R}(t)^{-1} B(t)$

Example: B constant $\Longrightarrow \mathcal{R}(t) = e^{tB}$

Explicit solution

Proposition 2

Consider:

- B, σ *satisfying the affine assumption,* X *solution of the affine equation*
- $\mathcal{R}$ *the resolvent of* B

Then almost surely, for all $t \in [0, \tau]$,

$$X_t = \mathcal{R}(t)\, X_0 + \mathcal{R}(t) \int_0^t \mathcal{R}(r)^{-1} \sigma(r)\, dW_r$$

Interpretation: Variation of constants

$\hookrightarrow$ Linear in X_0, plus a Wiener integral with deterministic integrand

Proof

Itô's formula: For $f(t,x) = \mathcal{R}(t)^{-1}x$, linear in x
$\hookrightarrow$ No second order term

$$\begin{aligned} d\left(\mathcal{R}(t)^{-1}X_t\right) &= \left[\tfrac{d}{dt}\,\mathcal{R}(t)^{-1}\right] X_t\,dt + \mathcal{R}(t)^{-1}\,dX_t \\ &= \mathcal{R}(t)^{-1}\left[-B(t)X_t\,dt + B(t)X_t\,dt + \sigma(t)\,dW_t\right] \\ &= \mathcal{R}(t)^{-1}\sigma(t)\,dW_t \end{aligned}$$

Conclusion: Integrate on $[0,t]$, then multiply by $\mathcal{R}(t)$

Gaussian vectors

Gaussian vector: $\theta \cdot G$ real Gaussian for all $\theta \in \mathbb{R}^d$, mean m, covariance Σ
$\hookrightarrow$ Affine images: $MG + c \sim \mathcal{N}(Mm + c,\, M\Sigma M^*)$

Density: If Σ positive definite, $G \sim \mathcal{N}(m, \Sigma)$ has density $g_\Sigma(y - m)$

$$g_\Sigma(z) = \frac{1}{(2\pi)^{d/2}\,(\det \Sigma)^{1/2}} \exp\left(-\tfrac{1}{2}\, z^* \Sigma^{-1} z\right)$$

Simulation: $Z \sim \mathcal{N}(0, \mathrm{Id})$, $\Sigma^{1/2}$ symmetric square root

$$m + \Sigma^{1/2} Z \sim \mathcal{N}(m, \Sigma)$$

Covariance matrix

Definition 3

Let $\mathcal{R}$ the resolvent of B and $a = \sigma\sigma^*$. Then:

- Two-parameter resolvent: $\mathcal{R}_{s,t} = \mathcal{R}(t)\,\mathcal{R}(s)^{-1}$, for $0 \leq s \leq t \leq \tau$
- Covariance matrix: for $t \in (0, \tau]$,

$$\Sigma(t) = \int_0^t \mathcal{R}_{r,t}\, a(r)\, \mathcal{R}_{r,t}^*\, dr$$

Lemma 4

For $t \in (0, \tau]$, the matrix $\Sigma(t)$ is symmetric and positive definite

Gaussian transition

Proposition 5

Consider:

- B, σ *satisfying the affine assumption,* X *solution of the affine equation*
- $\mathcal{R}$ *the resolvent,* $\Sigma(t)$ *the covariance matrix,* $t \in (0, \tau]$

Then:

$$\mathcal{L}\left(X_t \,|\, X_0 = x\right) = \mathcal{N}\left(\mathcal{R}(t)\, x,\, \Sigma(t)\right)$$

Transition density: Gaussian density $g_{\Sigma(t)}$ recentered

$$\widetilde{q}_{0,t}(x, y) = g_{\Sigma(t)}\left(y - \mathcal{R}(t)\, x\right)$$

Proof

Start at x: Explicit solution with $X_0 = x$, and $\mathcal{R}(t)\,\mathcal{R}(r)^{-1} = \mathcal{R}_{r,t}$

$$X_t = \mathcal{R}(t)\,x + \int_0^t \mathcal{R}_{r,t}\,\sigma(r)\,dW_r$$

Wiener integral:

- Deterministic integrand $\Longrightarrow$ centered Gaussian
- Covariance given by the Itô isometry

Conclusion:

$$X_t \sim \mathcal{N}\left(\mathcal{R}(t)\,x,\ \underbrace{\int_0^t \mathcal{R}_{r,t}\,\sigma(r)\sigma(r)^*\,\mathcal{R}_{r,t}^*\,dr}_{=\Sigma(t)}\right)$$

Exact noising

Proposition 6

Consider:

- B, σ *satisfying the affine assumption,* $x \in \mathbb{R}^n$, $t \in (0, \tau]$
- $Z \sim \mathcal{N}(0, \mathrm{Id})$ *standard Gaussian vector on* $\mathbb{R}^n$

Then:

$$\mathcal{R}(t)\, x + \Sigma(t)^{1/2} Z \sim \mathcal{L}\left(X_t \,|\, X_0 = x\right)$$

Exact noising: proof and comments

Proof: Gaussian simulation, with $m = \mathcal{R}(t)x$ and $\Sigma = \Sigma(t)$
$\hookrightarrow$ Law identified by the Gaussian transition proposition

The noising step is exact:

- One draw of n i.i.d. $\mathcal{N}(0,1)$ per data point and time t
- No time discretization, hence no error at step (i)
- Nonlinear drift: numerical scheme, and an extra error

Time reversal in the affine setting

Theorem 7

Consider:

- *Affine setting, $X_0 \sim \mu$ arbitrary, p_t density of X_t, $\varepsilon \in (0, \tau)$*

Then p is smooth, positive, solves Fokker–Planck, and there is a Wiener process $\widehat{W}$ such that on $[0, \tau - \varepsilon]$

$$d\widehat{X}_t = \left[-B(\tau - t)\widehat{X}_t + a(\tau - t)\nabla \log p_{\tau - t}(\widehat{X}_t)\right] dt + \sigma(\tau - t)\, d\widehat{W}_t$$

Proof (admitted): Explicit kernel, Kolmogorov equations by hand, Fisher bound

Variance preserving coefficients

Coefficients: $\beta : [0, \tau] \to (0, \infty)$ smooth, $\bar{\beta}_t = \int_0^t \beta(r)\, dr$

$$B(t) = -\tfrac{1}{2}\beta(t)\,\mathrm{Id}, \quad \sigma(t) = \sqrt{\beta(t)}\,\mathrm{Id}$$

Lemma 8

For the variance preserving coefficients,

$$\mathcal{R}(t) = e^{-\bar{\beta}_t/2}\,\mathrm{Id}, \quad a(t) = \beta(t)\,\mathrm{Id}, \quad \Sigma(t) = \left(1 - e^{-\bar{\beta}_t}\right)\mathrm{Id}$$

Hence $\mathcal{L}(X_t \,|\, X_0 = x) = \mathcal{N}\left(e^{-\bar{\beta}_t/2}x,\, (1 - e^{-\bar{\beta}_t})\,\mathrm{Id}\right)$

Variance preserving equation: two features

Variance preserved: μ centered, $\mathbf{Cov}(X_0) = \mathrm{Id} \Rightarrow \mathbf{Cov}(X_t) = \mathrm{Id}$

Relaxation: $\bar{\beta}_t \to \infty \Longrightarrow \mathcal{L}(X_t \,|\, X_0 = x) \to \mathcal{N}(0, \mathrm{Id})$ at rate $e^{-\bar{\beta}_t}$
$\hookrightarrow$ Justifies step (iii): start the reversed equation from $\mathcal{N}(0, \mathrm{Id})$

Exact noising: Data point x, $Z \sim \mathcal{N}(0, \mathrm{Id})$

$$e^{-\bar{\beta}_t/2}\, x + \left(1 - e^{-\bar{\beta}_t}\right)^{1/2} Z \sim \mathcal{L}\left(X_t \,|\, X_0 = x\right)$$

Outline

1. The generative algorithm
2. The affine forward equation
3. The score-matching identity
4. Denoising and training

The fitting problem

Object to estimate: True score, for $t \in (0, \tau]$ and $x \in \mathbb{R}^n$,

$$s_t^\star(x) = \nabla \log p_t(x)$$

Model: Parametric family $s_\theta(t, x)$, $\theta \in \Theta$, typically a network
$\hookrightarrow$ Natural mean-square loss at time t:

$$L_t(\theta) = \mathbf{E}\left[\left|s_\theta(t, X_t) - s_t^\star(X_t)\right|^2\right]$$

Problem: $s_t^\star$ unknown, so L_t cannot be computed

Score-matching identity

Theorem 9 (Hyvärinen)

Consider:

- X *forward diffusion as in Chapter 5,* $X_0 \sim \mu$, $t \in (0, \tau]$
- s_θ *bounded,* $\mathcal{C}^1$ *in* x *with bounded derivatives*

Then:

$$L_t(\theta) = \mathbf{E}\left[|s_\theta(t, X_t)|^2 + 2\,\nabla \cdot s_\theta(t, X_t)\right] + I(p_t)$$

Fisher information: $I(p_t) = \mathbf{E}\left[|s_t^\star(X_t)|^2\right]$, independent of θ

$\hookrightarrow$ Right-hand side free of $s^\star$, up to a constant

Proof

Expanding the square: Integral against p_t,

$$L_t(\theta) = \int_{\mathbb{R}^n} |s_\theta(t,x)|^2 \, p_t(x) \, dx - 2 \int_{\mathbb{R}^n} \langle s_\theta(t,x), \nabla p_t(x) \rangle \, dx + I(p_t)$$

$\hookrightarrow$ Used $p_t \, \nabla \log p_t = \nabla p_t$ in the cross term

Integration by parts: p_t, ∇p_t decay at infinity

$$\begin{aligned} \int_{\mathbb{R}^n} \langle s_\theta(t,x), \nabla p_t(x) \rangle \, dx &= - \int_{\mathbb{R}^n} \nabla \cdot s_\theta(t,x) \, p_t(x) \, dx \\ &= -\mathbf{E} \left[\nabla \cdot s_\theta(t, X_t) \right] \end{aligned}$$

Conclusion: Plug the second relation into the first

Outline

1. The generative algorithm
2. The affine forward equation
3. The score-matching identity
4. Denoising and training

Strategy

Drawback of score matching: $\nabla \cdot s_\theta = \sum_i \partial_{x_i} s_{\theta,i}$
$\hookrightarrow$ Autodiff: one network pass per coordinate, n in total

Remedy: Use the Gaussian transition of the affine equation
$\hookrightarrow$ Affine forward equation from now on

Aim: Explicit Gaussian vector Z_t, coupled to X_t, with

$$\mathbf{E}\left[-\Sigma(t)^{-1/2} Z_t \,|\, X_t\right] = s_t^\star(X_t)$$

Consequence: $-\Sigma(t)^{-1/2} Z_t$ as a regression target for $s_\theta(t, X_t)$

- No divergence of s_θ any more
  $\hookrightarrow$ Loss evaluated in one network pass, whatever n
- Same minimizers as L_t (Vincent)

The noise of the noising step

Definition 10

Let $\mathcal{R}$ the resolvent and $\Sigma(t)$ the covariance matrix. For $t \in (0, \tau]$, the noise at time t is

$$Z_t = \Sigma(t)^{-1/2} \left(X_t - \mathcal{R}(t) X_0 \right), \quad \text{i.e.} \quad X_t = \mathcal{R}(t) X_0 + \Sigma(t)^{1/2} Z_t$$

Lemma 11

For $t \in (0, \tau]$:

- $Z_t \sim \mathcal{N}(0, \mathrm{Id})$, *independent of* X_0
- (X_0, X_t) *has law* $\widetilde{q}_{0,t}(z, y)\, \mu(dz)\, dy$

Tweedie's identity

Conditional score: Gradient of the Gaussian kernel in y

$$\nabla_y \log \widetilde{q}_{0,t}(X_0, X_t) = -\Sigma(t)^{-1}\left(X_t - \mathcal{R}(t)X_0\right) = -\Sigma(t)^{-1/2} Z_t$$

Proposition 12

Consider the affine setting, $X_0 \sim \mu$ arbitrary, $t \in (0, \tau]$. Then:

$$s_t^\star(X_t) = \mathbf{E}\left[-\Sigma(t)^{-1/2} Z_t \mid X_t\right] \quad \textit{almost surely}$$

In words: True score = conditional mean of the conditional score

Proof

Same three steps as the Brownian case:

Step 1, differentiation: Under the integral in $p_t(y) = \int \widetilde{q}_{0,t}(z,y)\,\mu(dz)$,

$$\nabla p_t(y) = \int_{\mathbb{R}^n} -\Sigma(t)^{-1}\,(y - \mathcal{R}(t)z)\;\widetilde{q}_{0,t}(z,y)\,\mu(dz)$$

Step 2, Bayes rule: Law of X_0 given $X_t = y$ is $\widetilde{q}_{0,t}(z,y)\,\mu(dz)/p_t(y)$

Step 3, division by $p_t(y)$:

$$\nabla \log p_t(y) = \mathbf{E}\left[-\Sigma(t)^{-1}\,(X_t - \mathcal{R}(t)X_0) \mid X_t = y\right]$$

Denoising loss

Definition 13

Let Z_t the noise at time t and s_θ a parametric family. The denoising loss is

$$J_t(\theta) = \mathbf{E}\left[\left|s_\theta(t, X_t) + \Sigma(t)^{-1/2} Z_t\right|^2\right]$$

Interpretation:

- Regression of the target $-\Sigma(t)^{-1/2} Z_t$ on X_t
- Target known: it is the noise added at step (i)
- No derivative of s_θ

Denoising score matching

Theorem 14 (Vincent)

Consider:

- *Affine setting,* $X_0 \sim \mu$, $t \in (0, \tau]$
- $s_\theta(t, \cdot)$ *Borel, with* $\mathbf{E}[|s_\theta(t, X_t)|^2] < \infty$

Then:

$$L_t(\theta) = J_t(\theta) - tr\left(\Sigma(t)^{-1}\right) + I(p_t)$$

Consequence: L_t and J_t have the same minimizers over $\theta \in \Theta$

Proof

Notation: $A = s_\theta(t, X_t)$, $S = s_t^\star(X_t)$, $T = -\Sigma(t)^{-1/2} Z_t$
$\hookrightarrow L_t(\theta) = \mathbf{E}[|A - S|^2]$ and $J_t(\theta) = \mathbf{E}[|A - T|^2]$

Orthogonality: Tweedie gives $\mathbf{E}[T \,|\, X_t] = S$
$\hookrightarrow \mathbf{E}[\langle U, T - S \rangle] = 0$ for $U \in L^2$, $\sigma(X_t)$-measurable

Pythagoras: With $U = A - S$, then $U = S$, and $T \sim \mathcal{N}(0, \Sigma(t)^{-1})$,

$$J_t(\theta) = L_t(\theta) + \mathbf{E}\left[|S - T|^2\right] = L_t(\theta) + \mathbf{E}\left[|T|^2\right] - \mathbf{E}\left[|S|^2\right]$$
$$= L_t(\theta) + \mathrm{tr}\left(\Sigma(t)^{-1}\right) - I(p_t)$$

Weighted loss and noise prediction

All times at once: Weight $w \geq 0$ on $[\varepsilon, \tau]$, under growth and moment conditions

$$\int_{\varepsilon}^{\tau} w(t)\, L_t(\theta)\, dt = \int_{\varepsilon}^{\tau} w(t)\, J_t(\theta)\, dt - C_w$$

$\hookrightarrow$ $C_w \geq 0$ independent of θ

Noise prediction: $s_\theta(t, x) = -\Sigma(t)^{-1/2} \epsilon_\theta(t, x)$
$\hookrightarrow$ Network predicts the noise, optimal predictor by Tweedie:

$$\epsilon^\star(t, X_t) = \mathbf{E}\left[Z_t \mid X_t\right]$$

Noise prediction: variance preserving case

VP coefficients: $\mathcal{R}(t) = e^{-\bar{\beta}_t/2}\,\mathrm{Id}$, $\Sigma(t) = (1 - e^{-\bar{\beta}_t})\,\mathrm{Id}$

$$J_t(\theta) = \frac{1}{1 - e^{-\bar{\beta}_t}}\,\mathbf{E}\left[\left|Z_t - \epsilon_\theta\left(t,\, e^{-\bar{\beta}_t/2}X_0 + (1 - e^{-\bar{\beta}_t})^{1/2}Z_t\right)\right|^2\right]$$

Weight $w(t) = 1 - e^{-\bar{\beta}_t}$: Prefactor disappears

- Plain mean-square error between noise and prediction
- Loss of denoising diffusion probabilistic models
- Constant $C_w \leq n(\tau - \varepsilon)$, uniformly in ε

Feedforward neural network

Definition 15

Let:

- Depth $m \geq 1$, widths $d_0 = n + 1$, $d_1, \ldots, d_{m-1}$, $d_m = n$
- Activation $\rho : \mathbb{R} \to \mathbb{R}$ continuous, acting coordinatewise
- Parameter $\theta = (M_1, c_1, \ldots, M_m, c_m)$, $M_k \in \mathbb{R}^{d_k \times d_{k-1}}$, $c_k \in \mathbb{R}^{d_k}$

The network is $s_\theta(t, x) = h_m$, with $h_0 = (t, x)$ and

$$h_k = \rho\left(M_k h_{k-1} + c_k\right), \ 1 \leq k \leq m-1, \qquad h_m = M_m h_{m-1} + c_m$$

Unbiased estimator of the weighted loss

Lemma 16

Consider, all independent:

- $x \sim \mu$, T *uniform on* $[\varepsilon, \tau]$, $Z \sim \mathcal{N}(0, \mathrm{Id})$
- *Noised sample* $X = \mathcal{R}(T)\, x + \Sigma(T)^{1/2} Z$

Then:

$$(\tau - \varepsilon)\, \mathbf{E}\left[w(T) \left| s_\theta(T, X) + \Sigma(T)^{-1/2} Z \right|^2 \right] = \int_\varepsilon^\tau w(t)\, J_t(\theta)\, dt$$

Batch loss: P i.i.d. copies $(x_j, T_j, Z^{(j)})$, unbiased by the lemma

$$\widehat{J}_P(\theta) = \frac{\tau - \varepsilon}{P} \sum_{j=1}^{P} w(T_j) \left| s_\theta(T_j, X^{(j)}) + \Sigma(T_j)^{-1/2} Z^{(j)} \right|^2$$

Stochastic gradient descent with minibatches

Data: Network, weight w, batch size P, step sizes λ_k, $k_{\max}$ iterations

Algorithm:

1. Initial parameter θ_0, at random
2. At step k: fresh indices I_j, times T_j, noises $Z^{(j)}$, $j \leq P$
3. Noised samples $X^{(j)} = \mathcal{R}(T_j)x_{I_j} + \Sigma(T_j)^{1/2}Z^{(j)}$, batch loss $\widehat{J}_P$
4. Gradient step, with $\nabla_\theta \widehat{J}_P$ computed by autodiff:

$$\theta_{k+1} = \theta_k - \lambda_k \nabla_\theta \widehat{J}_P(\theta_k)$$

5. Return $\theta_{k_{\max}}$

Variance preserving case

Choices: Noise predictor ϵ_θ, weight $w(t) = 1 - e^{-\bar{\beta}_t}$
$\hookrightarrow$ Batch loss:

$$\widehat{J}_P(\theta) = \frac{\tau - \varepsilon}{P} \sum_{j=1}^{P} \left| Z^{(j)} - \epsilon_\theta \left(T_j,\ e^{-\bar{\beta}_{T_j}/2} x_{I_j} + (1 - e^{-\bar{\beta}_{T_j}})^{1/2} Z^{(j)} \right) \right|^2$$

Literature: Training algorithm of Ho, Jain, Abbeel (DDPM)
$\hookrightarrow$ Times drawn from a finite grid instead of $[\varepsilon, \tau]$

Convergence of SGD: Not discussed (non-convex loss)
$\hookrightarrow$ Chapter 7 takes the fitted s_θ as given

Summary

Affine forward equation:

- Gaussian transitions, exact noising in one draw
- Time reversal of Chapter 5 still available

Training phase:

- Score matching: L_t rewritten without $s^\star$, up to a constant
  $\hookrightarrow$ Computable from samples of X_t, but needs $\nabla \cdot s_\theta$
- Tweedie and denoising: explicit regression on the noise
- Minibatch SGD on an unbiased estimator

Next (Chapter 7): Sampling phase and total error