Sampling and error analysis

Samy Tindel

Purdue University

Stochastic calculus and Machine Learning - MA 595

PURDUE UNIVERSITY

Outline

1. The sampling phase
2. Relative entropy
3. The change of measure estimate
4. Cutoff, discretization and total error

Outline

1. The sampling phase
2. Relative entropy
3. The change of measure estimate
4. Cutoff, discretization and total error

Where we stand

Forward equation (Chapter 6): $X_0 \sim \mu$, affine coefficients

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

Training phase (Chapter 6): Network $s_\theta(t, x) \approx s_t^\star(x) = \nabla \log p_t(x)$
$\hookrightarrow$ Parameter θ fixed from now on

This chapter:

- Steps iii–iv of the algorithm: initialization and sampling
- Error between the law of the output and the data law μ

The ideal reversed equation

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

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

$\hookrightarrow$ $a = \sigma\sigma^*$, cutoff $\varepsilon \in (0, \tau)$ fixed

Out of reach of the user:

- Initial value $\widehat{X}_0 = X_\tau$, whose law p_τ depends on μ
- True score $s^\star$

Sampling phase: Replace both, then discretize

The initial law

Law of X_τ: Gaussian transition $\Longrightarrow$ mixture of Gaussians

$$p_\tau(y) = \int_{\mathbb{R}^n} g_{\Sigma(\tau)}\left(y - \mathcal{R}(\tau)z\right)\mu(dz)$$

$\hookrightarrow$ Data enter only through the means $\mathcal{R}(\tau)z$

Reference law: Means set to 0, sampled exactly

$$\gamma_\tau = \mathcal{N}\left(0, \Sigma(\tau)\right), \qquad \Sigma(\tau)^{1/2} Z \sim \gamma_\tau \quad \text{for } Z \sim \mathcal{N}(0, \mathrm{Id})$$

VP case: $\gamma_\tau = \mathcal{N}(0, (1 - e^{-\bar{\beta}_\tau})\mathrm{Id})$, close to $\mathcal{N}(0, \mathrm{Id})$ used in practice
$\hookrightarrow$ Below: an arbitrary initial law ν

The fitted score

Assumption:

i. s_θ continuous on $[\varepsilon, \tau] \times \mathbb{R}^n$, Lipschitz in x with constant Λ_θ
ii. Bounded error, namely

$$K_\theta = \sup \{|s_\theta(t,x) - s_t^\star(x)|;\ t \in [\varepsilon, \tau],\ x \in \mathbb{R}^n\} < \infty$$

Item i: Property of the network alone (next frame)

Item ii: Not a bound on s_θ. For $\mu = \delta_0$, $p_t = g_{\Sigma(t)}$ and
$\hookrightarrow$ $s_t^\star(x) = -\Sigma(t)^{-1}x$, linear growth to be reproduced

Regularity of a network

Lemma 1

Consider:

- s_θ *network of Chapter 6, with matrices* $M_1, \dots, M_m$
- *Activation* ρ *Lipschitz, with constant* κ_ρ

Then s_θ *is continuous, and for* $s, t \in \mathbb{R}$, $x, y \in \mathbb{R}^n$,

$$|s_\theta(t,x) - s_\theta(s,y)| \leq \Lambda_\theta \left(|t-s| + |x-y| \right), \quad \Lambda_\theta = \kappa_\rho^{m-1} \prod_{k=1}^{m} |M_k|$$

Consequences: Item i of the assumption holds

$\hookrightarrow$ Lipschitz in time: used for the discretization

Continuous-time sampler

Definition 2

Let:

- Drift $\widehat{b}_\theta(t,x) = -B(\tau - t)\,x + a(\tau - t)\,s_\theta(\tau - t, x)$
- $\widehat{W}$ Wiener process for $(\widehat{\mathcal{F}}_t)$, $\eta \in \widehat{\mathcal{F}}_0$ with law ν

The continuous-time sampler solves, for $t \in [0, \tau - \varepsilon]$,

$$\widehat{X}^\theta_t = \eta + \int_0^t \widehat{b}_\theta(r, \widehat{X}^\theta_r)\,dr + \int_0^t \sigma(\tau - r)\,d\widehat{W}_r$$

Its output is $\widehat{X}^\theta_{\tau-\varepsilon}$, with law ρ^θ

Continuous-time sampler: comments

Drift error: Linear parts cancel against the reversed drift $\widehat{b}$

$$\widehat{b}_\theta(t,x) - \widehat{b}(t,x) = a(\tau - t)\left[s_\theta(\tau - t, x) - s^\star_{\tau - t}(x)\right]$$

Well posed: Additive noise, drift Lipschitz with $L = \|B\|_\infty + \|a\|_\infty \Lambda_\theta$
$\hookrightarrow$ Existence and uniqueness of Chapter 3, random initial condition

Ideal case: $s_\theta = s^\star$ and $\nu = p_\tau \Longrightarrow \widehat{X}^\theta = \widehat{X}$
$\hookrightarrow$ Output X_ε, with law p_ε: exact up to the cutoff

Euler–Maruyama sampler

Grid: $0 = t_0 < \cdots < t_K = \tau - \varepsilon$, steps $h_k = t_{k+1} - t_k$, mesh $\delta = \max_k h_k$

Definition 3

The Euler–Maruyama sampler is $\widehat{X}_{t_0}^{\theta,\delta} = \eta$ and, for $k \leq K-1$,

$$\widehat{X}_{t_{k+1}}^{\theta,\delta} = \widehat{X}_{t_k}^{\theta,\delta} + \widehat{b}_\theta\left(t_k, \widehat{X}_{t_k}^{\theta,\delta}\right) h_k + \sigma(\tau - t_k)\left(\widehat{W}_{t_{k+1}} - \widehat{W}_{t_k}\right)$$

Its output is $\widehat{X}_{\tau-\varepsilon}^{\theta,\delta} = \widehat{X}_{t_K}^{\theta,\delta}$, with law $\rho^{\theta,\delta}$

Sampling phase in practice

Increments: $\widehat{W}_{t_{k+1}} - \widehat{W}_{t_k} = \sqrt{h_k}\, Z_{k+1}$
$\hookrightarrow$ $Z_1, \ldots, Z_K$ i.i.d. $\mathcal{N}(0, \mathrm{Id})$, independent of η

VP case: $\widehat{b}_\theta(t,x) = \beta(\tau - t)\left[\frac{1}{2}x + s_\theta(\tau - t, x)\right]$, hence

$$\widehat{X}^{\theta,\delta}_{t_{k+1}} = \widehat{X}^{\theta,\delta}_{t_k} + \beta(\tau - t_k)\left[\tfrac{1}{2}\widehat{X}^{\theta,\delta}_{t_k} + s_\theta\left(\tau - t_k, \widehat{X}^{\theta,\delta}_{t_k}\right)\right] h_k + \sqrt{\beta(\tau - t_k)\, h_k}\; Z_{k+1}$$

Cost: One network evaluation per step
$\hookrightarrow$ DDPM sampler: exponential integrator variant (Nakano)

Error split

Output laws: $\rho^\theta = \mathcal{L}(\widehat{X}^\theta_{\tau-\varepsilon})$ and $\rho^{\theta,\delta} = \mathcal{L}(\widehat{X}^{\theta,\delta}_{\tau-\varepsilon})$
$\hookrightarrow$ Independent of the probability space (checked later for ρ^θ)

Triangle inequality: For any distance d on probability measures,

$$d\left(\mu, \rho^{\theta,\delta}\right) \leq d\left(\mu, p_\varepsilon\right) + d\left(p_\varepsilon, \rho^\theta\right) + d\left(\rho^\theta, \rho^{\theta,\delta}\right)$$

Aim: Bound each term in terms of ε, τ, ν, θ, δ

Three terms, three tools

Sources of error:

- $d(\mu, p_\varepsilon)$: cutoff, forward equation only
  $\hookrightarrow$ Compare X_0 and X_ε: coupling
- $d(p_\varepsilon, \rho^\theta)$: initialization and score (goals iii–iv)
  $\hookrightarrow$ Same diffusion coefficient: change of measure
- $d(\rho^\theta, \rho^{\theta,\delta})$: discretization (goal v)
  $\hookrightarrow$ Same η and $\widehat{W}$: coupling

Distances: Total variation for the second term
$\hookrightarrow$ A weaker distance for the total error

Outline

1. The sampling phase
2. Relative entropy
3. The change of measure estimate
4. Cutoff, discretization and total error

Total variation and relative entropy

Definition 4

Let P, Q probability measures on $(E, \mathcal{E})$. Then:

- Total variation: $d_{\mathrm{TV}}(P, Q) = \sup\{|P(A) - Q(A)|;\ A \in \mathcal{E}\}$
- Relative entropy: $+\infty$ if $P \not\ll Q$, otherwise

$$\mathrm{KL}(P \,\|\, Q) = \int_E \frac{dP}{dQ} \log\left(\frac{dP}{dQ}\right) dQ$$

Properties:

- $\mathrm{KL}(P \,\|\, Q) \in [0, \infty]$ by Jensen, not symmetric
- Densities: $\mathrm{KL}(p \,\|\, q) = \int_{\{p>0\}} p \log(p/q)$, if $q > 0$

Data processing inequality

Lemma 5

Consider:

- P, Q *probability measures on* $(E, \mathcal{E})$
- $F : E \to E'$ *measurable, image measures* $P \circ F^{-1}$, $Q \circ F^{-1}$

Then:

$$\mathrm{KL}\left(P \circ F^{-1} \,\|\, Q \circ F^{-1}\right) \leq \mathrm{KL}(P \,\|\, Q)$$

Use: Compare two processes, conclude on their values at one time

Proof (notes): Density of $P \circ F^{-1}$ is $\mathbf{E}_Q[dP/dQ \,|\, F]$
$\hookrightarrow$ Conditional Jensen for $x \mapsto x \log x$

Pinsker's inequality

Lemma 6 (Pinsker)

Let P, Q probability measures on $(E, \mathcal{E})$. Then:

$$d_{\mathrm{TV}}(P, Q) \leq \left(\tfrac{1}{2}\,\mathrm{KL}(P \,\|\, Q)\right)^{1/2}$$

Use: Relative entropy is what Girsanov computes
$\hookrightarrow$ Pinsker converts it into total variation

Proof (1)

Optimal set: $\mathrm{KL}(P \,\|\, Q) < \infty$, $f = dP/dQ$, $A = \{f \geq 1\}$. For $C \in \mathcal{E}$,

$$P(C) - Q(C) = \int_C (f-1)\, dQ \leq \int_A (f-1)\, dQ = P(A) - Q(A)$$

$\hookrightarrow$ Same with C^c, hence $d_{\mathrm{TV}}(P, Q) = P(A) - Q(A)$

Reduction: Data processing with $F = \mathbf{1}_A$, $a = P(A)$, $c = Q(A)$

$$\mathrm{KL}(P \,\|\, Q) \geq k(c) := a \log \frac{a}{c} + (1-a) \log \frac{1-a}{1-c}$$

Proof (2)

Claim: $k(c) \geq 2(a-c)^2$, for $c \in (0,1)$ (cases $c \in \{0,1\}$ trivial)

Derivative: With $h(c) = k(c) - 2(a-c)^2$,

$$h'(c) = \frac{c-a}{c(1-c)} + 4(a-c) = (c-a)\left[\frac{1}{c(1-c)} - 4\right]$$

$\hookrightarrow$ Bracket ≥ 0, since $c(1-c) \leq \frac{1}{4}$

Conclusion: h decreases then increases, $h(a) = 0$, so $h \geq 0$, and

$$\mathrm{KL}(P \,\|\, Q) \geq 2\,(P(A) - Q(A))^2 = 2\,d_{\mathrm{TV}}(P,Q)^2$$

Outline

1. The sampling phase
2. Relative entropy
3. The change of measure estimate
4. Cutoff, discretization and total error

Girsanov's theorem

Theorem 7 (Girsanov)

Consider:

- W *Wiener process for* $(\mathcal{G}_t)_{t\in[0,T]}$ *under* $\mathbf{P}$
- u *progressively measurable,* $|u_t| \leq C$

Set $Z_T = \exp\left(\int_0^T u_r \cdot dW_r - \frac{1}{2}\int_0^T |u_r|^2\, dr\right)$. *Then* $\mathbf{E}[Z_T \,|\, \mathcal{G}_0] = 1$, *and under* $d\mathbf{Q} = Z_T\, d\mathbf{P}$,

$$\widetilde{W}_t = W_t - \int_0^t u_r\, dr \quad \text{is a Wiener process for } (\mathcal{G}_t)$$

Proof (admitted): Karatzas–Shreve, Novikov criterion and Lévy's characterization

The law of the sampler

Lemma 8

Consider:

- *Affine setting, fitted score as in the assumption,* $T = \tau - \varepsilon$
- $\mathcal{C}_T = \mathcal{C}([0, T]; \mathbb{R}^n)$ *with the supremum norm*

Then there is a continuous $\Phi : \mathbb{R}^n \times \mathcal{C}_T \to \mathcal{C}_T$*, depending on* B, σ, s_θ *only, with*

$$\widehat{X}^\theta = \Phi\left(\eta, \widehat{W}\right) \quad \textit{almost surely, on any probability space}$$

Consequence: ρ^θ depends on ν only

$\hookrightarrow$ The sampler may be built on a space chosen for convenience

Proof

Noise term: Itô's formula for $(r, x) \mapsto f(r)\, x$, with $f(r) = \sigma(\tau - r)$

$$\int_0^t \sigma(\tau - r)\, d\widehat{W}_r = f(t)\, \widehat{W}_t - \int_0^t f'(r)\, \widehat{W}_r\, dr =: \Psi(\widehat{W})_t$$

$\hookrightarrow \Psi : \mathcal{C}_T \to \mathcal{C}_T$ linear and continuous

Random ODE: $Y = \widehat{X}^\theta - \Psi(\widehat{W})$ solves, path by path,

$$Y_t = \eta + \int_0^t \widehat{b}_\theta \left(r, Y_r + \Psi(\widehat{W})_r\right) dr$$

$\hookrightarrow$ Solution $\Gamma(\eta, \Psi(\widehat{W}))$, continuous by Gronwall. Set $\Phi = \Psi + \Gamma(\cdot, \Psi)$

Score and initialization error

Theorem 9

Consider:

- *Affine setting, fitted score as in the assumption*
- *Initial law ν with a positive density*

Then:

$$\mathrm{KL}\left(p_\varepsilon \,\|\, \rho^\theta\right) \leq \mathrm{KL}\left(p_\tau \,\|\, \nu\right) + \frac{1}{2}\int_\varepsilon^\tau \mathbf{E}\left[\left|\sigma(t)^*\left(s_\theta(t,\cdot) - s_t^\star\right)(X_t)\right|^2\right] dt$$

Key feature: Expectation along the forward X, not the sampler
$\hookrightarrow$ A quantity of the training phase

Proof (1)

Strategy: Build the sampler on the space of the ideal reversed process $\hookrightarrow \widehat{X}$, $\widehat{W}$, $(\widehat{\mathcal{F}}_t)$ of the reversal theorem, $T = \tau - \varepsilon$

Drift correction: Bounded by $\|\sigma\|_\infty K_\theta$, Girsanov applies

$$u_t = \sigma(\tau - t)^* \left[s_\theta(\tau - t, \widehat{X}_t) - s^\star_{\tau - t}(\widehat{X}_t) \right]$$

$\hookrightarrow$ Drift error along $\widehat{X}$: $\widehat{b}_\theta(t, \widehat{X}_t) - \widehat{b}(t, \widehat{X}_t) = \sigma(\tau - t)\, u_t$

New measure: Z_T Girsanov density of u, and initial correction

$$d\mathbf{Q} = D\, Z_T\, d\mathbf{P}, \qquad D = \frac{\nu(\widehat{X}_0)}{p_\tau(\widehat{X}_0)}$$

Proof (2)

Probability: $D \in \widehat{\mathcal{F}}_0$, so $\mathbf{E}[DZ_T] = \mathbf{E}[D\,\mathbf{E}[Z_T|\widehat{\mathcal{F}}_0]] = \mathbf{E}[D] = 1$

Initial law under $\mathbf{Q}$: For g bounded, conditioning on $\widehat{\mathcal{F}}_0$,

$$\mathbf{E}_{\mathbf{Q}}\left[g(\widehat{X}_0)\right] = \mathbf{E}_{\mathbf{P}}\left[g(\widehat{X}_0)\,D\right] = \int_{\mathbb{R}^n} g(y)\,\frac{\nu(y)}{p_\tau(y)}\,p_\tau(y)\,dy = \int_{\mathbb{R}^n} g\,\nu$$

Noise under $\mathbf{Q}$: $\widetilde{W} = \widehat{W} - \int_0^{\cdot} u_r\,dr$

- Wiener process under $Z_T\,d\mathbf{P}$, by Girsanov
- Still Wiener under $\mathbf{Q}$: $D \in \widehat{\mathcal{F}}_0$ factors out of each increment

Proof (3)

Rewriting the reversed equation: $\widehat{b} = \widehat{b}_\theta - \sigma u$ and $d\widehat{W} = d\widetilde{W} + u\,dr$

$$\widehat{X}_t = \widehat{X}_0 + \int_0^t \widehat{b}_\theta(r, \widehat{X}_r)\,dr + \int_0^t \sigma(\tau - r)\,d\widetilde{W}_r$$

$\hookrightarrow$ Under $\mathbf{Q}$: sampler from ν, so $\mathcal{L}_{\mathbf{Q}}(\widehat{X}_T) = \rho^\theta$ by the lemma

$\hookrightarrow$ Under $\mathbf{P}$: $\widehat{X}_T = X_\varepsilon$, so $\mathcal{L}_{\mathbf{P}}(\widehat{X}_T) = p_\varepsilon$

Data processing: With $F = \widehat{X}_T$,

$$\mathrm{KL}\left(p_\varepsilon \,\|\, \rho^\theta\right) \leq \mathrm{KL}\left(\mathbf{P} \,\|\, \mathbf{Q}\right)$$

Proof (4)

Entropy of the change of measure: $d\mathbf{P}/d\mathbf{Q} = (DZ_T)^{-1}$

$$\begin{aligned}
\mathrm{KL}(\mathbf{P} \,\|\, \mathbf{Q}) &= \mathbf{E}_{\mathbf{P}}[-\log D] + \mathbf{E}_{\mathbf{P}}[-\log Z_T] \\
&= \mathrm{KL}(p_\tau \,\|\, \nu) + \mathbf{E}_{\mathbf{P}}\left[-\int_0^T u_r \cdot d\widehat{W}_r + \frac{1}{2}\int_0^T |u_r|^2 \, dr\right] \\
&= \mathrm{KL}(p_\tau \,\|\, \nu) + \frac{1}{2}\int_0^T \mathbf{E}_{\mathbf{P}}\left[|u_r|^2\right] dr
\end{aligned}$$

$\hookrightarrow$ Itô integral of a bounded integrand: mean zero

Conclusion: $|u_r|^2$ evaluated along $\widehat{X}_r = X_{\tau - r}$
$\hookrightarrow$ Change of variable $t = \tau - r$ gives the bound of the theorem

Score and initialization error in total variation

Pinsker: With $|\sigma^* v|^2 \leq |a|\,|v|^2$ in the theorem,

$$d_{\mathrm{TV}}\left(p_\varepsilon, \rho^\theta\right) \leq \left(\tfrac{1}{2}\,\mathrm{KL}\left(p_\tau \,\|\, \nu\right) + \tfrac{1}{4}\int_\varepsilon^\tau |a(t)|\, L_t(\theta)\, dt\right)^{1/2}$$

Score term: $L_t(\theta) = \mathbf{E}[|s_\theta(t, X_t) - s_t^\star(X_t)|^2]$, loss of Chapter 6

- Weighted loss, weight $|a(t)|$: $\beta(t)$ for VP, not $1 - e^{-\bar{\beta}_t}$
- K_θ absent from the bound, only used to apply Girsanov

Initialization term: Depends on μ, the forward equation and ν only

Initialization from the reference law

Proposition 10

Consider:

- *Affine setting, reference law* $\gamma_\tau = \mathcal{N}(0, \Sigma(\tau))$
- μ *with second moment* $m_2(\mu) = \int_{\mathbb{R}^n} |z|^2 \, \mu(dz) < \infty$

Then:

$$\begin{aligned} \mathrm{KL}\left(p_\tau \,\|\, \gamma_\tau\right) &\leq \frac{1}{2} \int_{\mathbb{R}^n} \left|\Sigma(\tau)^{-1/2} \mathcal{R}(\tau) z\right|^2 \mu(dz) \\ &\leq \frac{1}{2} \left|\Sigma(\tau)^{-1/2} \mathcal{R}(\tau)\right|^2 m_2(\mu) \end{aligned}$$

Interpretation: $p_\tau \approx \gamma_\tau$ when the means $\mathcal{R}(\tau)z$ are small
$\hookrightarrow$ Compared with the spread $\Sigma(\tau)^{1/2}$

Proof

Mixture: $p_\tau = \int p_z \, \mu(dz)$ with p_z density of $\mathcal{N}(\mathcal{R}(\tau)z, \Sigma(\tau))$
$\hookrightarrow$ Jensen for $x \mapsto x \log x$, pointwise in y, then Tonelli:

$$\mathrm{KL}\left(p_\tau \,\|\, \gamma_\tau\right) \leq \int_{\mathbb{R}^n} \mathrm{KL}\left(p_z \,\|\, \gamma_\tau\right) \mu(dz)$$

Gaussian computation: $G \sim \mathcal{N}(m, \Sigma)$, log-ratio ℓ of the densities
$\hookrightarrow$ Same normalization, so $\ell(y) = m^* \Sigma^{-1} y - \frac{1}{2} m^* \Sigma^{-1} m$

$$\mathrm{KL}\left(\mathcal{N}(m, \Sigma) \,\|\, \mathcal{N}(0, \Sigma)\right) = \mathbf{E}\left[\ell(G)\right] = \frac{1}{2} \left|\Sigma^{-1/2} m\right|^2$$

$\hookrightarrow$ Apply with $m = \mathcal{R}(\tau)z$, $\Sigma = \Sigma(\tau)$

Variance preserving case

Corollary 11

Consider the VP equation, $\nu = \mathcal{N}(0, \mathrm{Id})$, $m_2(\mu) < \infty$, $x_\tau = e^{-\bar{\beta}_\tau}$. *Then:*

$$\mathrm{KL}\left(p_\tau \,\|\, \mathcal{N}(0, \mathrm{Id})\right) \leq \frac{x_\tau}{2}\, m_2(\mu) + \frac{n\, x_\tau^2}{4(1 - x_\tau)}$$

Consequence: Weight $|a(t)| = \beta(t)$ in the theorem

$$d_{\mathrm{TV}}\left(p_\varepsilon, \rho^\theta\right) \leq \left(\frac{x_\tau}{4}\, m_2(\mu) + \frac{n\, x_\tau^2}{8(1 - x_\tau)} + \frac{1}{4}\int_\varepsilon^\tau \beta(t)\, L_t(\theta)\, dt\right)^{1/2}$$

$\hookrightarrow$ Exponentially small in $\bar{\beta}_\tau$, plus a weighted training loss

Outline

1. The sampling phase
2. Relative entropy
3. The change of measure estimate
4. Cutoff, discretization and total error

Bounded-Lipschitz distance

Cutoff in total variation: μ empirical law, carried by a null set A
$\hookrightarrow \mu(A) - p_\varepsilon(A) = 1$, so $d_{\mathrm{TV}}(\mu, p_\varepsilon) = 1$ for every $\varepsilon > 0$

Definition 12

Let BL_1 be the set of $f : \mathbb{R}^n \to \mathbb{R}$ with $|f| \leq 1$ and $|f(x) - f(y)| \leq |x - y|$. The bounded-Lipschitz distance is

$$d_{\mathrm{BL}}(P, Q) = \sup \left\{ \left| \int_{\mathbb{R}^n} f \, dP - \int_{\mathbb{R}^n} f \, dQ \right| ; \, f \in \mathrm{BL}_1 \right\}$$

Bounded-Lipschitz distance: properties

Lemma 13

Let P, Q probability measures on $\mathbb{R}^n$. Then:

i. *d_{BL} is a distance, and $d_{\mathrm{BL}}(P,Q) \leq 2\, d_{\mathrm{TV}}(P,Q)$*
ii. *Coupling: if $X \sim P$ and $Y \sim Q$ live on the same space,*

$$d_{\mathrm{BL}}(P,Q) \leq \mathbf{E}\left[|X-Y|\right] \leq \mathbf{E}\left[|X-Y|^2\right]^{1/2}$$

Item ii: $|\mathbf{E}[f(X)] - \mathbf{E}[f(Y)]| \leq \mathbf{E}[|X-Y|]$ for $f \in \mathrm{BL}_1$

Weak convergence: $d_{\mathrm{BL}}(P_k, P) \to 0$ iff $P_k \to P$ weakly (Dudley)

The cutoff error

Proposition 14

Consider:

- *Affine setting,* $\|B\|_\infty, \|a\|_\infty$ *suprema of operator norms on* $[0, \tau]$
- μ *with second moment* $m_2(\mu) < \infty$

Then:

$$d_{\mathrm{BL}}(\mu, p_\varepsilon) \leq \left(|\mathcal{R}(\varepsilon) - \mathrm{Id}|^2 \, m_2(\mu) + \operatorname{tr} \Sigma(\varepsilon) \right)^{1/2}$$
$$\leq \varepsilon^{1/2} \, e^{\varepsilon \|B\|_\infty} \left(\varepsilon \, \|B\|_\infty^2 \, m_2(\mu) + n \, \|a\|_\infty \right)^{1/2}$$

VP case: Bound $\left(\frac{1}{4} \bar{\beta}_\varepsilon^2 \, m_2(\mu) + n \, \bar{\beta}_\varepsilon \right)^{1/2}$

$\hookrightarrow$ Dominant term $(n\bar{\beta}_\varepsilon)^{1/2}$: size of the noise added up to ε

Proof

Coupling: Explicit solution, $\mathcal{R}(\varepsilon)G_\varepsilon \sim \mathcal{N}(0, \Sigma(\varepsilon))$ independent of X_0

$$X_\varepsilon - X_0 = (\mathcal{R}(\varepsilon) - \mathrm{Id})\, X_0 + \mathcal{R}(\varepsilon)\, G_\varepsilon$$

$\hookrightarrow$ Cross term centered, $\mathbf{E}[|\mathcal{R}(\varepsilon)G_\varepsilon|^2] = \operatorname{tr}\Sigma(\varepsilon)$
$\hookrightarrow$ Coupling property of d_{BL} gives the first bound

Resolvent: $\mathcal{R}_{r,t} = \mathrm{Id} + \int_r^t B(u)\mathcal{R}_{r,u}\, du$, Gronwall

$$|\mathcal{R}_{r,t}| \leq e^{(t-r)\|B\|_\infty}, \qquad |\mathcal{R}(\varepsilon) - \mathrm{Id}| \leq \varepsilon\, \|B\|_\infty\, e^{\varepsilon\|B\|_\infty}$$

$\hookrightarrow \operatorname{tr}\Sigma(\varepsilon) \leq n\,|\Sigma(\varepsilon)| \leq n\,\varepsilon\,\|a\|_\infty\, e^{2\varepsilon\|B\|_\infty}$

Interpolated Euler–Maruyama sampler

Left grid point: $\kappa(0) = 0$ and $\kappa(r) = t_k$ for $r \in (t_k, t_{k+1}]$

Interpolation: Same recipe between grid points, continuous in t

$$\widehat{X}_t^{\theta,\delta} = \eta + \int_0^t \widehat{b}_\theta\left(\kappa(r), \widehat{X}_{\kappa(r)}^{\theta,\delta}\right) dr + \int_0^t \sigma(\tau - \kappa(r))\, d\widehat{W}_r$$

$\hookrightarrow$ Sampler equation with coefficients frozen at the grid points

Coupling: Same η and same $\widehat{W}$ as the continuous-time sampler

Discretization error

Theorem 15

Consider:

- *Affine setting, fitted score as in the assumption*
- s_θ *Lipschitz in time, constant* Λ_θ *(true for a network)*
- ν *with finite second moment, mesh* $\delta \leq 1$

Then, for a constant C_{disc} *independent of* δ,

$$d_{\text{BL}}\left(\rho^{\theta}, \rho^{\theta,\delta}\right) \leq C_{\text{disc}}\, \delta^{1/2}$$

Constant: $C_{\text{disc}} = C_3\, e^{LT}(1 + n + M_\nu)^{1/2}$

$\hookrightarrow$ M_ν: bound on $\mathbf{E}[|\widehat{X}_{t_k}^{\theta,\delta}|^2]$, C_3 depends on B, σ, θ, T

Proof (1)

Noise terms: Itô isometry, σ Lipschitz and $r - \kappa(r) \leq \delta$

$$N_t = \int_0^t [\sigma(\tau - r) - \sigma(\tau - \kappa(r))]\, d\widehat{W}_r, \qquad \mathbf{E}\left[|N_t|^2\right] \leq nT \left\|\sigma'\right\|_\infty^2 \delta^2$$

Drift terms: $F = \widehat{X}^\theta - \widehat{X}^{\theta,\delta} - N$ is a dr-integral of

$$\widehat{b}_\theta\left(r, \widehat{X}_r^\theta\right) - \widehat{b}_\theta\left(\kappa(r), \widehat{X}_{\kappa(r)}^{\theta,\delta}\right)$$

$\hookrightarrow$ Space: $\widehat{X}_r^\theta \to \widehat{X}_r^{\theta,\delta} \to \widehat{X}_{\kappa(r)}^{\theta,\delta}$, Lipschitz with constant L
$\hookrightarrow$ Time: $r \to \kappa(r)$, drift Lipschitz in time, $\leq K_2\delta(1 + |x|)$

Proof (2)

Gronwall, path by path: $|F_t| \leq \Xi + L \int_0^t |F_r|\, dr$, with

$$\Xi = \int_0^T \left(L|N_r| + L \left| \widehat{X}_r^{\theta,\delta} - \widehat{X}_{\kappa(r)}^{\theta,\delta} \right| + K_2\, \delta \left(1 + |\widehat{X}_{\kappa(r)}^{\theta,\delta}| \right) \right) dr$$

$\hookrightarrow$ Hence $|F_T| \leq \Xi\, e^{LT}$

Moments of the scheme: By induction on the steps

$$\mathbf{E}\left[|\widehat{X}_{t_k}^{\theta,\delta}|^2 \right] \leq M_\nu, \qquad \mathbf{E}\left[\left| \widehat{X}_r^{\theta,\delta} - \widehat{X}_{\kappa(r)}^{\theta,\delta} \right|^2 \right] \leq c_2\, \delta$$

$\hookrightarrow$ $\mathbf{E}[\Xi^2] \leq c_3\, \delta$, so $\mathbf{E}[|\widehat{X}_T^{\theta} - \widehat{X}_T^{\theta,\delta}|^2] \leq C_{\text{disc}}^2\, \delta$: coupling

Discretization error: comments

Rate $\delta^{1/2}$: Wiener increment over one step, of size $h_k^{1/2}$

- Order δ possible if $\nabla_x s_\theta$ Lipschitz (tanh yes, ReLU no)
- Uses orthogonal martingale increments

Constant: Grows like e^{LT}, with L linear in Λ_θ

- $\mu = \delta_0$: true score $-\Sigma(t)^{-1}x$, blows up as $t \to 0$
- Expect Λ_θ of order $|\Sigma(\varepsilon)^{-1}|$: large near the cutoff

Total error of the algorithm

Theorem 16

Consider:

- *Hypotheses of the discretization theorem,* $m_2(\mu) < \infty$
- ν *with a positive density*

Then:

$$d_{\mathrm{BL}}\left(\mu, \rho^{\theta,\delta}\right) \leq \varepsilon^{1/2}\, e^{\varepsilon \|B\|_\infty} \left(\varepsilon \|B\|_\infty^2\, m_2(\mu) + n \|a\|_\infty\right)^{1/2} + \left(2\,\mathrm{KL}\left(p_\tau \,\|\, \nu\right) + \int_\varepsilon^\tau |a(t)|\, L_t(\theta)\, dt\right)^{1/2} + C_{\mathrm{disc}}\, \delta^{1/2}$$

Total error: proof and reading

Proof: Error split with $d = d_{\mathrm{BL}}$

- Cutoff term: cutoff proposition
- Middle term: $d_{\mathrm{BL}} \leq 2\, d_{\mathrm{TV}}$, then score and initialization theorem
- Last term: discretization theorem

One parameter per term:

- Cutoff ε: order $\varepsilon^{1/2}$, constant growing like $n^{1/2}$
- Horizon τ and training: initialization and weighted loss
- Mesh δ: order $\delta^{1/2}$

Total error: variance preserving case

Corollary 17

Consider the VP equation, $\nu = \mathcal{N}(0, \mathrm{Id})$, $x_\tau = e^{-\bar{\beta}_\tau}$, and the hypotheses of the total error theorem. Then:

$$d_{\mathrm{BL}}\left(\mu, \rho^{\theta,\delta}\right) \leq \left(\tfrac{1}{4}\,\bar{\beta}_\varepsilon^2\, m_2(\mu) + n\,\bar{\beta}_\varepsilon\right)^{1/2} + C_{\mathrm{disc}}\,\delta^{1/2} + \left(x_\tau\, m_2(\mu) + \frac{n\, x_\tau^2}{2(1-x_\tau)} + \int_\varepsilon^\tau \beta(t)\, L_t(\theta)\, dt\right)^{1/2}$$

Reading: Choose ε small, $\bar{\beta}_\tau$ large, train, then $\delta \to 0$

$\hookrightarrow$ C_{disc} depends on ε and θ, so the mesh is chosen last

Summary

Sampling phase:

- Reversed equation with the fitted score, from a Gaussian law
- Euler–Maruyama scheme, one network call per step

Error analysis:

- Score and initialization: Girsanov, relative entropy, Pinsker
- Cutoff and discretization: couplings
- Total error in the bounded-Lipschitz distance