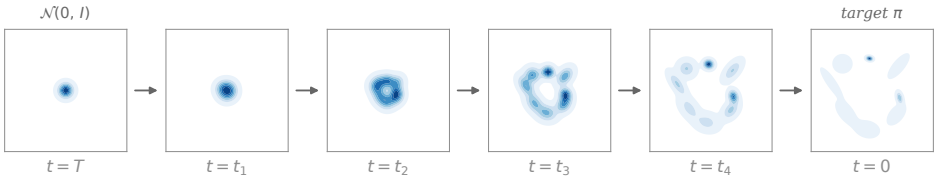
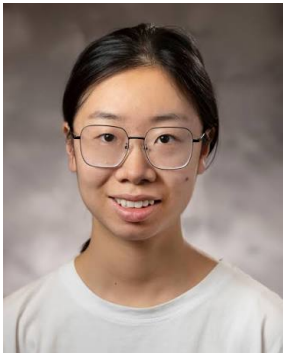
# Markov Chain Monte Carlo with Diffusion Paths

Jun Yang

Department of Mathematical Sciences
University of Copenhagen



## Joint work with



**Han Chen**
Duke University

**Sifan Liu**
Duke University

Markov Chain Monte Carlo with Diffusion Paths

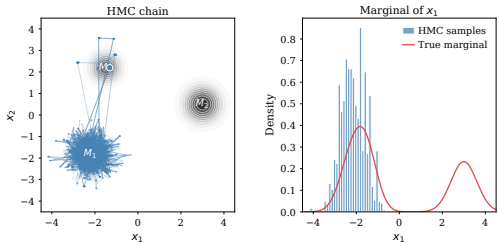
arXiv:2607.11631

# Markov chain Monte Carlo

- Target $p$ on $\mathbb{R}^d$, known only up to a normalizing constant.
- MCMC builds a chain with stationary distribution $p$.

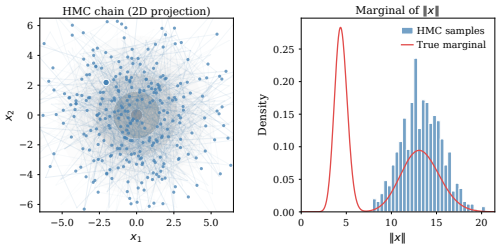
Local moves get stuck in isolated regions.

$$p(\mathbf{x}) = \tfrac{1}{3}\sum_{k=1}^{3} \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$




separated modes

$$p(\mathbf{x}) = \tfrac{1}{2}\sum_{k=1}^{2} \mathcal{N}(\mathbf{x} \mid \mathbf{0}, \sigma_k^2 I_d)$$




two centered Gaussians

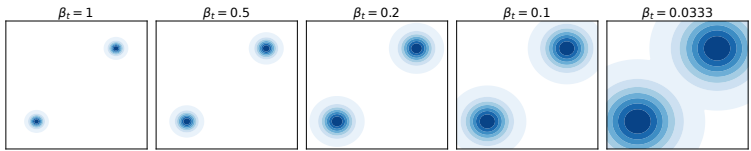
Difficulty Level 1: discover the regions. Level 2: move across the regions

# Interpolation: bridging to a simpler distribution

Flatten the target by tempering:

$$p_\beta(\mathbf{x}) \;\propto\; p(\mathbf{x})^\beta, \qquad \beta \in [0, 1].$$

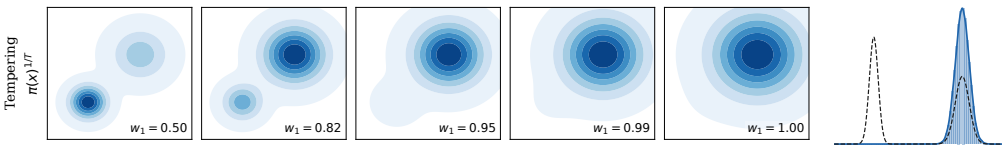
- Parallel tempering [Swendsen & Wang, 1986]
- Simulated tempering [Geyer & Thompson, 1995]
- Tempered transition [Neal, 1996]



# Tempering can distort relative weights

Prone to mode collapse when the isolated regions differ in scale:

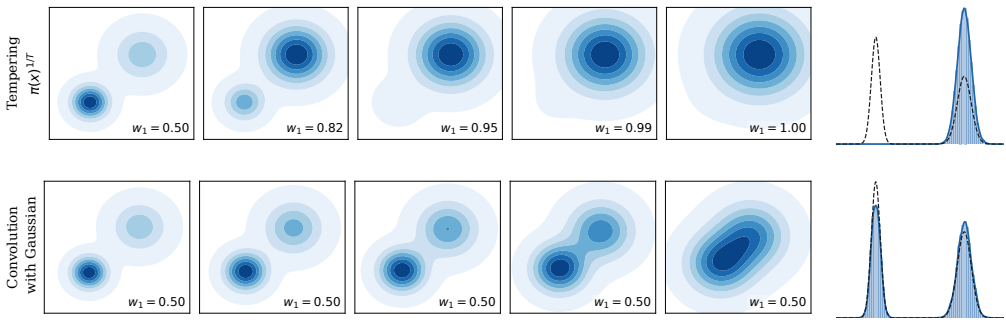
$$p(\mathbf{x}) = \tfrac{1}{2}\,\mathcal{N}(\mathbf{x} \mid \mathbf{1}_{20},\, 0.2 \cdot I_{20}) + \tfrac{1}{2}\,\mathcal{N}(\mathbf{x} \mid -\mathbf{1}_{20},\, 0.5 \cdot I_{20})$$



# Tempering can distort relative weights

Prone to mode collapse when the isolated regions differ in scale:

$$p(\mathbf{x}) = \tfrac{1}{2}\,\mathcal{N}(\mathbf{x} \mid \mathbf{1}_{20},\, 0.2 \cdot I_{20}) + \tfrac{1}{2}\,\mathcal{N}(\mathbf{x} \mid -\mathbf{1}_{20},\, 0.5 \cdot I_{20})$$




- Convolution preserves the mixture weights at every noise level.

# Interpolation in the sample space

Interpolate in the *sample space* rather than in temperature:

$$X \sim p, \qquad Z \sim \mathcal{N}(0, I_d),$$

$$X_t = \sqrt{\alpha_t}\, X + \sqrt{1 - \alpha_t}\, Z, \qquad t \geq 0.$$

$\mathrm{Law}(X_t)$ preserves the mixture weights of $p$.

# Diffusion path

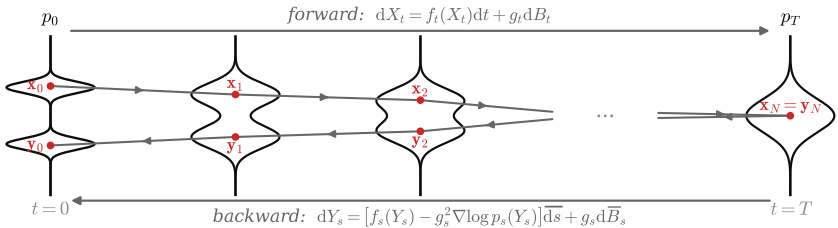
Ornstein–Uhlenbeck process — the continuous-time limit of repeated Gaussian convolution:

$$\mathrm{d}X_t = -\tfrac{1}{2} X_t \,\mathrm{d}t + \mathrm{d}B_t, \qquad X_0 \sim p.$$

Its *time reversal*, with $p_t = \mathrm{Law}(X_t)$ and score $s_t = \nabla \log p_t$:

$$\mathrm{d}Y_t = \left[-\tfrac{1}{2} Y_t - s_t(Y_t)\right] \overline{\mathrm{d}t} + \mathrm{d}\overline{B}_t, \qquad Y_T \sim p_T.$$

[Anderson, 1982 · Song et al., 2021]



# Where does the score come from?

**Generative modeling**

- Samples from $p$ are given.
- Learn $s_t$ by score matching on data.

[Song et al., 2021]

**Bayesian computation**

- Only the unnormalized density.
- No samples to regress on.

A rich and active literature learns $s_t$ in the Bayesian setting:

- stochastic optimal control / path-space variational inference,
- simulation-free regression via the target score identity,
- diffusion combined with sequential Monte Carlo, PDE-based methods.

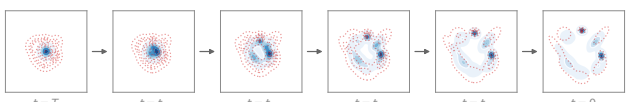
[Zhang & Chen, 2022 · Vargas et al., 2023 · De Bortoli et al., 2024 · Noble et al., 2025]

In practice we only have $\hat{s}_t \approx \nabla \log p_t$, how to design MCMC which is $p$ invariant?

# A simple sampler — and three sources of error

**Simple sampler:** draw $\mathbf{y}_N \sim \mathcal{N}(0, I_d)$, then run the backward pass with $\hat{s}_t$ and step size $h$.

1. **Truncation** ($T < \infty$): $\mathcal{N}(0, I_d) \neq p_T$.
2. **Score** ($\hat{s}_t \neq s_t$).
3. **Discretization** ($h > 0$).



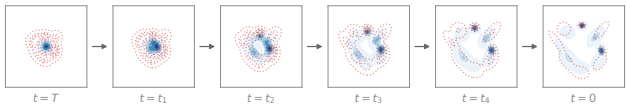

a wrong initialization never recovers

Reducing one typically increases another.

# A simple sampler — and three sources of error

**Simple sampler:** draw $\mathbf{y}_N \sim \mathcal{N}(0, I_d)$, then run the backward pass with $\hat{s}_t$ and step size $h$.

1. **Truncation** ($T < \infty$): $\mathcal{N}(0, I_d) \neq p_T$.
2. **Score** ($\hat{s}_t \neq s_t$).
3. **Discretization** ($h > 0$).




a wrong initialization never recovers

Reducing one typically increases another.

Can we remove all three — exactly?

# Removing the truncation error

Do not guess $p_T$ — reach it by running the forward process first.

With exact forward and backward diffusions, this defines the **ideal transition kernel** $P$ on $\mathbf{x}_0$:

$$X_T \mid X_0 \sim \mathcal{N}\Big(\sqrt{\alpha}\, X_0,\, (1-\alpha) I_d\Big), \qquad \alpha = e^{-T},$$
$$X_0' \mid X_T \sim p_{X_0 \mid X_T}(\cdot \mid X_T).$$

- $P$ is equivalent to **proximal sampler** [Lee et al.'21, Chen et al.'22, Wibisono'26]
- $P$ is the marginal of a **two-block data-augmentation Gibbs sampler** on $(X_0, X_T)$.

So we can borrow the theory from Gibbs sampling.

# Spectral gap of $P$

**Lemma**

$$\mathrm{Gap}(P) = 1 - \rho_{\max}(X_0, X_T)^2,$$

*where the maximal correlation is*

$$\rho_{\max}(X_0, X_T) := \sup_{f \in L^2(p_0),\ g \in L^2(p_T)} \mathrm{Corr}\big(f(X_0),\, g(X_T)\big).$$

*[Liu et al., 1994 · Bradley, 2005]*

# Spectral gap of $P$

**Lemma**

$$\mathrm{Gap}(P) = 1 - \rho_{\max}(X_0, X_T)^2,$$

*where the maximal correlation is*

$$\rho_{\max}(X_0, X_T) := \sup_{f \in L^2(p_0),\ g \in L^2(p_T)} \mathrm{Corr}\big(f(X_0),\, g(X_T)\big).$$

*[Liu et al., 1994 · Bradley, 2005]*

The less correlated $X_0$ and $X_T$, the faster the chain mixes.

- Large $T$: $X_T$ nearly independent of $X_0$ $\Rightarrow$ gap close to $1$.
- Small $T$: $X_T \approx X_0$ $\Rightarrow$ gap close to $0$.

- For a mixture of *nice* (Poincaré) components,

  the gap depends on neither the mixture weights nor the relative scales.

# How large must $T$ be?

The unequal-variance mixture that defeats tempering:

$$p = \tfrac{1}{2}\,\mathcal{N}\big(b\mathbf{1}_d,\, \sigma_+^2 I_d\big) + \tfrac{1}{2}\,\mathcal{N}\big(-b\mathbf{1}_d,\, \sigma_-^2 I_d\big), \qquad \sigma_+ > \sigma_-.$$

**Corollary**

*Recall $\alpha = e^{-T}$. Fix $b, \sigma_\pm$ and set $\alpha = \gamma/d$ for a fixed $\gamma > 0$. Then*

$$\liminf_{d\to\infty} \operatorname{Gap}(P) \;\geq\; e^{-\gamma b^2}.$$

$T = \Theta(\log d)$ suffices for a dimension-free gap.

# How large must $T$ be?

The unequal-variance mixture that defeats tempering:

$$p = \tfrac{1}{2}\,\mathcal{N}\big(b\mathbf{1}_d,\, \sigma_+^2 I_d\big) + \tfrac{1}{2}\,\mathcal{N}\big(-b\mathbf{1}_d,\, \sigma_-^2 I_d\big), \qquad \sigma_+ > \sigma_-.$$

**Corollary**

*Recall $\alpha = e^{-T}$. Fix $b, \sigma_\pm$ and set $\alpha = \gamma/d$ for a fixed $\gamma > 0$. Then*

$$\liminf_{d\to\infty} \operatorname{Gap}(P) \;\geq\; e^{-\gamma b^2}.$$

$T = \Theta(\log d)$ suffices for a dimension-free gap.

- By contrast, mixing time of tempering is $\Omega(\exp(d))$ for any temperature ladder [Woodard et al.'09].

# Approximate the ideal transition kernel

Given an approximate reverse drift $\hat{b}_t(\mathbf{y}) = \frac{1}{2}\mathbf{y} + \hat{s}_t(\mathbf{y})$ and step size $h = T/N$:

**Start** from the current state $\mathbf{x}_0$.

**Forward pass** · for $k = 0, \dots, N-1$

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \tfrac{1}{2}\mathbf{x}_k h + \sqrt{h}\,\varepsilon_k, \qquad \varepsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d).$$

**Backward pass** · start at $\mathbf{y}_N = \mathbf{x}_N$; for $k = N-1, \dots, 0$

$$\mathbf{y}_k \leftarrow \mathbf{y}_{k+1} + \hat{b}_{t_k}(\mathbf{y}_{k+1})\, h + \sqrt{h}\,\zeta_k, \qquad \zeta_k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d).$$

**Arrive** at $\mathbf{y}_0$.

NOT $p$ invariant: score error and discretization error

# Removing the other two errors

Score and discretization errors are still there. Ideally we would Metropolize:

$$\alpha(\mathbf{x}, \mathbf{x}') = \min\left\{1,\ \frac{p_t(\mathbf{x}')\, q_t(\mathbf{x} \mid \mathbf{x}')}{p_t(\mathbf{x})\, q_t(\mathbf{x}' \mid \mathbf{x})}\right\}$$

- But the intermediate marginals $p_t$ are intractable.

  (Recall $\hat{s}_t \approx \nabla \log p_t$ is never exact)

- MH is not directly applicable.

# Removing the other two errors

Score and discretization errors are still there. Ideally we would Metropolize:

$$\alpha(\mathbf{x}, \mathbf{x}') = \min\left\{1,\ \frac{p_t(\mathbf{x}')\, q_t(\mathbf{x} \mid \mathbf{x}')}{p_t(\mathbf{x})\, q_t(\mathbf{x}' \mid \mathbf{x})}\right\}$$

- But the intermediate marginals $p_t$ are intractable.

  (Recall $\hat{s}_t \approx \nabla \log p_t$ is never exact)

- MH is not directly applicable.

**Our approach:**

Metropolis–Hastings in *path space*

# Augmented state and path reversal

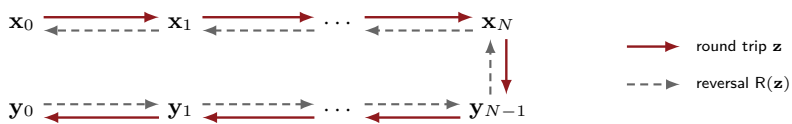
- Augmented state $\mathbf{z} := (\mathbf{x}_0, \dots, \mathbf{x}_N, \mathbf{y}_{N-1}, \dots, \mathbf{y}_0)$, with joint density

$$\Pi(\mathbf{z}) = p(\mathbf{x}_0) \prod_{k=0}^{N-1} \vec{p}_k(\mathbf{x}_{k+1} \mid \mathbf{x}_k) \prod_{k=0}^{N-1} \overleftarrow{p}_k(\mathbf{y}_k \mid \mathbf{y}_{k+1}),$$

$$\vec{p}_k(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \varphi\big(\mathbf{x}_{k+1};\, \mathbf{x}_k - \tfrac{1}{2}\mathbf{x}_k h,\, hI_d\big),$$
$$\overleftarrow{p}_k(\mathbf{y}_k \mid \mathbf{y}_{k+1}) = \varphi\big(\mathbf{y}_k;\, \mathbf{y}_{k+1} + \hat{b}_{t_k}(\mathbf{y}_{k+1})h,\, hI_d\big).$$

- Proposal: reverse the round-trip path:

$$\mathsf{R}(\mathbf{x}_0, \dots, \mathbf{x}_N, \mathbf{y}_{N-1}, \dots, \mathbf{y}_0) := (\mathbf{y}_0, \dots, \mathbf{y}_{N-1}, \mathbf{x}_N, \mathbf{x}_{N-1}, \dots, \mathbf{x}_0).$$



# Metropolis-Adjusted Diffusion Path Sampler (MAD-Path)

- So a standard MH correction applies: accept $\mathbf{y}_0$ with probability

$$\frac{\Pi(\mathsf{R}\mathbf{z})}{\Pi(\mathbf{z})} \wedge 1 = \frac{p(\mathbf{y}_0)}{p(\mathbf{x}_0)} \cdot \frac{\prod_{k=0}^{N-1} \overleftarrow{p}_k(\mathbf{x}_k \mid \mathbf{x}_{k+1}) \cdot \prod_{k=0}^{N-1} \vec{p}_k(\mathbf{y}_{k+1} \mid \mathbf{y}_k)}{\prod_{k=0}^{N-1} \vec{p}_k(\mathbf{x}_{k+1} \mid \mathbf{x}_k) \cdot \prod_{k=0}^{N-1} \overleftarrow{p}_k(\mathbf{y}_k \mid \mathbf{y}_{k+1})} \wedge 1.$$

**Theorem**

*MAD-Path leaves the target $p$ invariant — regardless of the score error and discretization error.*

- Structural Analogy to HMC:

  Augmented space → sample Momentum → leapfrog proposal → MH in the augmented space

- Similar to *tempered transition* (Neal, 1996) but using diffusion path instead of tempering

# Optimal scaling and practical guidance

For fixed $T$, let $N = T/h$ and $h = \ell \cdot d^{-1}$

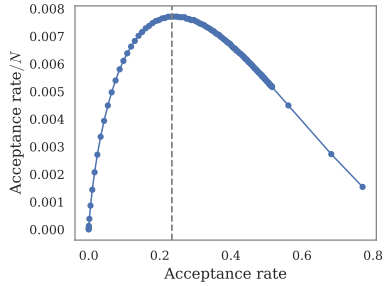
Write $\Delta$ for the log MH ratio, then under certain conditions

$$\Delta \Rightarrow \mathcal{N}\Big( -\tfrac{\ell \mathcal{D}(T)}{2},\, \ell \mathcal{D}(T) \Big),$$

where $\mathcal{D}(T)$ measures the discretization error accumulated over $[0, T]$.

Balancing acceptance against the cost $N$ gives

optimal acceptance rate $\approx 0.234$.



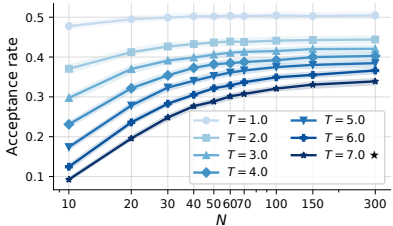

$\mathcal{N}(10 \cdot \mathbf{1}_d, 3^2 I_d)$, $d = 500$, $T = 2$

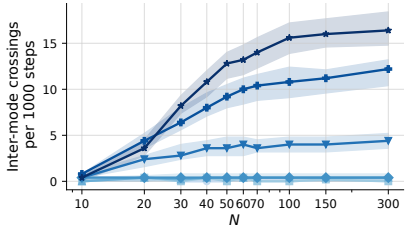
**Tuning recipe.**

1. Choose $T$ large enough for the isolated regions to overlap.
2. Given $T$, increase $N$ until the acceptance rate reaches $\approx 0.234$.
3. If acceptance saturates below $0.234$, the score error dominates. It is an integral over $[0, T]$, so it grows with $T$ — lower $T$ and retune.

$$\mathcal{E}(T) := \frac{1}{2} \int_0^T \Big\{ \mathbb{E}[\|\hat{s}_t(X_t) - s_t(X_t)\|^2 + \|\hat{s}_t(\hat{Y}_t) - s_t(\hat{Y}_t)\|^2] \Big\} \, \mathrm{d}t$$

# Effects of $N$ and $T$




- Acceptance rate increases with $N$ and decreases with $T$ — and saturates in $N$ at a ceiling.
- Inter-mode crossings improve with $T$.
- Trade-off: Larger $T$ improves the crossing between modes, but lowers the acceptance ceiling.

# Learning the score: we reuse existing methods

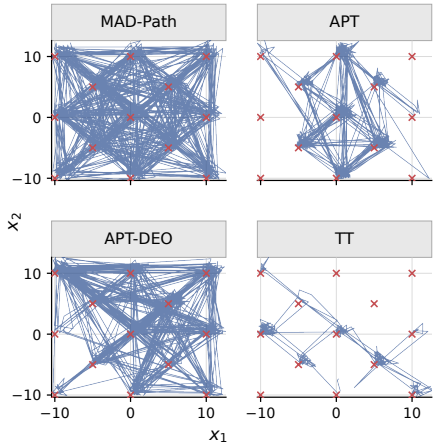
Nothing new here — MAD-Path is **modular** in $\hat{s}_t$.

- **Stochastic optimal control** / path-space variational inference [Zhang & Chen'22 · Vargas et al.'23]
  - ↪ Schrödinger bridges, Langevin preconditioning [Richter & Berner, 2024]
  - ↪ target score identity: simulation-free regression [De Bortoli et al., 2024]
- **Diffusion with SMC**: learned drifts, annealed resampling [Doucet et al.'22 · Akhound-Sadegh et al.'25]
- **PDE-based**: Fokker–Planck via physics-informed networks [Sun et al., 2024]
- **Reference distribution**: a fitted Gaussian mixture mitigates mode collapse [Noble et al., 2025]

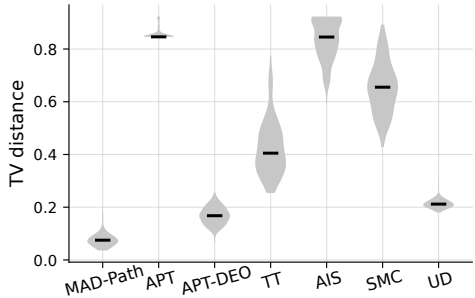
- Any future advance in score learning plugs straight in.
- A better score buys higher acceptance, therefore a larger $T$.

# Traces and mode weight estimation

Trace plots of the first two coordinates




Mode weight estimation




[APT: Miasojedow et al., 2013 · APT-DEO: Okabe, 2001 · TT: Neal, 1996 · AIS: Neal, 2001 · SMC: Del Moral et al., 2006 · UD: Noble et al., 2025]

- MAD-Path crosses between modes far more often, and its mode-weight estimates converge much faster

## Conclusions

- MAD-Path removes all three errors: for any time horizon $T$, estimated score $\hat{s}_t$ and step size $h$.
- Guidance for tuning: tuned to $\approx 0.234$.
- MAD-Path is efficient if $T$ is large enough: the integrated $L_2$ score error is small enough.

**Future directions:**

- Score-learning objectives aligned with MAD-Path

$$\text{integrated score error:} \quad \mathcal{E}(T) = \mathrm{KL}(\mathbb{P}_Y \| \mathbb{P}_{\hat{Y}}) + \mathrm{KL}(\mathbb{P}_{\hat{Y}} \| \mathbb{P}_Y)$$

- Adaptive MCMC: adaptively tuning of $T$ and $h$, and updating $\hat{s}_t$.
- Mixing-time analysis for MAD-Path itself.

***Thank you!***

Han Chen, Sifan Liu, JY, *Markov Chain Monte Carlo with Diffusion Paths*, `arXiv:2607.11631`