Stochastic Optimal Control Matching

Carles Domingo-Enrich, Jiequn Han, Brandon Amos, Joan Bruna, Ricky T. Q. Chen

Thomas Mousseau

August 26, 2025

Overview

Evolution of Generative Models

2020 **DDPM:** Denoising Diffusion Probabilistic Models interpret generation as reversing a discrete noise-adding process, learning to denoise at each step. They produced high-quality samples but required thousands of slow sampling steps.

2021 **Score-based Models:** Score-based generative models extended diffusion to continuous-time SDEs, learning the score function ($\nabla_x \log p_t(x)$) to reverse a stochastic diffusion process. This unified diffusion with stochastic control, allowed probability flow ODEs, and sped up sampling.

2023 **Flow Matching:** Flow matching views generation as learning a deterministic ODE vector field that directly transports a simple distribution (e.g., Gaussian) to data. This removed stochasticity and significantly improved efficiency compared to diffusion/score methods.

SOC as the Foundation of Generative Models

The Core Challenge: Unnormalized Densities

Generative models must sample from complex distributions $p_{\text{data}}(x) = \frac{1}{Z}\tilde{p}_{\text{data}}(x)$ where the normalization constant $Z = \int \tilde{p}_{\text{data}}(x)dx$ is intractable to compute. This intractability arises from the curse of dimensionality when integrating over high-dimensional spaces.

SOC Connection

Key Insight:
Transform tractable distributions (Gaussian) to complex target distributions through optimal control policies.

This bridges the gap between:

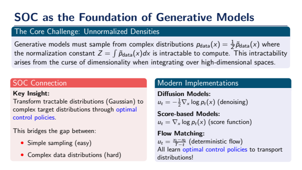
- Simple sampling (easy)
- Complex data distributions (hard)

Modern Implementations

Diffusion Models:
$u_t = -\frac{1}{2}\nabla_x \log p_t(x)$ (denoising)

Score-based Models:
$u_t = \nabla_x \log p_t(x)$ (score function)

Flow Matching:
$u_t = \frac{x_1 - x_0}{T-t}$ (deterministic flow)
All learn optimal control policies to transport distributions!



Unnormalized Densities: The fundamental challenge in generative modeling is sampling from distributions $p(x) = \frac{1}{Z}e^{-E(x)}$ where Z is unknown. SOC provides the mathematical framework to construct sampling procedures.

Historical Context: From Langevin dynamics to modern diffusion models, all major breakthroughs in generative modeling can be understood through the lens of stochastic optimal control theory.

What is a Stochastic Control Problem?

Control-Affine Stochastic Differential Equation

The general form of a controlled stochastic process:

$$dX_t^u = \big(b(X_t^u, t) + \sigma(t)u(X_t^u, t)\big)dt + \sqrt{\lambda}\sigma(t)dB_t \tag{1}$$

State Process: $X_t^u \in \mathbb{R}^d$ (system state under control u at time t)

Drift Term: $b(X_t^u, t) \in \mathbb{R}^d$ (natural evolution of the system)

Control Term: $\sigma(t)u(X_t^u, t) \in \mathbb{R}^d$ (how control influences the system)

Diffusion Coefficient: $\sigma(t) \in \mathbb{R}^{d\times d}$ (volatility matrix)

Noise Process: $B_t \in \mathbb{R}^d$ (Brownian motion, external randomness)

Noise Intensity: $\lambda > 0$ (controls the strength of stochastic perturbations)

What is a Stochastic Control Problem?

Control-Affine Stochastic Differential Equation

The general form of a controlled stochastic process:

$$dX_t^u = \left(b(X_t^u, t) + \sigma(t)u(X_t^u, t)\right)dt + \sqrt{\lambda}\sigma(t)dB_t \tag{1}$$

State Process: $X_t^u \in \mathbb{R}^d$ (system state under control u at time t)

Drift Term: $b(X_t^u, t) \in \mathbb{R}^d$ (natural evolution of the system)

Control Term: $\sigma(t)u(X_t^u, t) \in \mathbb{R}^d$ (how control influences the system)

Diffusion Coefficient: $\sigma(t) \in \mathbb{R}^{d\times d}$ (volatility matrix)

Noise Process: $B_t \in \mathbb{R}^d$ (Brownian motion, external randomness)

Noise Intensity: $\lambda > 0$ (controls the strength of stochastic perturbations)

Control-Affine Structure: The "control-affine" property means the control u enters linearly (affinely) in the drift term. This is the most general practical form for controlled SDEs, encompassing most applications in finance, robotics, and machine learning.

Je dois aussi expliquer l'importance que le steering input ainsi que le noise sont tous les 2 multipliés par sigma(t) ce qui signifie que ...

The Optimal Control Objective

Cost Function to Minimize

Find the optimal control policy $u^* \in \mathcal{U}$ that minimizes:

$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u(X_t^u, t)\|^2 + f(X_t^u, t)\right) dt + g(X_T^u)\right] \tag{2}$$

Control Effort: $\frac{1}{2}\|u(X_t^u, t)\|^2$ (penalizes large control actions)

Running Cost: $f(X_t^u, t)$ (ongoing cost during the process evolution)

Terminal Cost: $g(X_T^u)$ (final cost based on end state at time T)

Control Space: $\mathcal{U}$ (set of admissible control policies)

The Optimal Control Objective

Cost Function to Minimize

Find the optimal control policy $u^* \in \mathcal{U}$ that minimizes:

$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u(X_t^u, t)\|^2 + f(X_t^u, t)\right) dt + g(X_T^u)\right] \quad (2)$$

Control Effort: $\frac{1}{2}\|u(X_t^u, t)\|^2$ (penalizes large control actions)

Running Cost: $f(X_t^u, t)$ (ongoing cost during the process evolution)

Terminal Cost: $g(X_T^u)$ (final cost based on end state at time T)

Control Space: $\mathcal{U}$ (set of admissible control policies)

SOC in Practice

Aspect	Steering Actuator	Diffusion Models	Flow Models
State X_t	Vehicle position, velocity	Noisy data sample	Clean data sample
Control u_t	Steering angle, throttle	Denoising direction	Flow velocity field
Dynamics Source	Newtonian mechanics (vehicle dynamics)	Forward noising process	No natural drift (learnable)
Drift $b(X_t, t)$	Kinematic equations	Predetermined schedule	$b = 0$ (control learns drift)
Control Goal	Reach target safely	Reverse noise process	Transport distributions
Noise $\sqrt{\lambda}\sigma dB_t$	Road disturbances	Brownian motion	Optional stochasticity

SOC in Practice

Aspect	Steering Actuator	Diffusion Models	Flow Models
State X_t	Vehicle position, velocity	Noisy data sample	Clean data sample
Control u_t	Steering angle, throttle	Denoising direction	Flow velocity field
Dynamics Source	Newtonian mechanics (vehicle dynamics)	Forward noising process	No natural drift (learnable)
Drift $b(X_t, t)$	Kinematic equations	Predetermined schedule	$b = 0$ (control learns drift)
Control Goal	Reach target safely	Reverse noise process	Transport distributions
Noise $\sqrt{\lambda}\sigma dB_t$	Road disturbances	Brownian motion	Optional stochasticity

Steering Actuator: The dynamics come from well-established physics - Newton's laws, kinematics, and vehicle dynamics. The control optimizes safety and efficiency.

Diffusion Models: The drift is determined by the forward noising process (e.g., $\beta(t)$), and control learns to reverse this predetermined corruption.

Flow Models: No natural drift exists - the control u_t directly becomes the drift term, learning the entire velocity field that transports distributions.

Applications of SOC

Key ML Applications of SOC

- **Reward fine-tuning of diffusion and flow models:** Optimizing generation quality using reward signals
- **Conditional sampling on diffusion and flow models:** Steering generation towards specific conditions or constraints
- **Sampling from unnormalized densities:** Efficiently drawing samples from complex, intractable distributions
- **Importance sampling of rare events in SDEs:** Computing probabilities of low-probability but critical events

Key Insight

The prevalence of SOC formulations in modern ML motivates the need for more efficient and stable solving methods.

Applications of SOC

Key ML Applications of SOC

- **Reward fine-tuning of diffusion and flow models:** Optimizing generation quality using reward signals
- **Conditional sampling on diffusion and flow models:** Steering generation towards specific conditions or constraints
- **Sampling from unnormalized densities:** Efficiently drawing samples from complex, intractable distributions
- **Importance sampling of rare events in SDEs:** Computing probabilities of low-probability but critical events

Key Insight

The prevalence of SOC formulations in modern ML motivates the need for more efficient and stable solving methods.

Je dois etre capable de vraiment bien expliquer les applications du SOC dans le ML moderne pour pouvoir demontrer l'importance d'avoir pris ce papier

Solve SOC Problems in Low Dimension

Solve SOC Problems in High Dimension

Similar Trend in Generative Modeling

SOCM in Context: Optimization Landscapes

Task	Non-convex	Least Squares
Generative Modeling	Maximum Likelihood CNFs	Diffusion models and Flow Matching
Stochastic Optimal Control	Adjoint Methods	**Stochastic Optimal Control Matching**

SOCM Framework: From Classical SOC to SOCM

Dynamics (Same for Both)

$$dX_t^v = (b(X_t^v, t) + \sigma(t)v(X_t^v, t))dt + \sqrt{\lambda}\sigma(t)dB_t, \text{ with } X_0^v \sim p_0 \tag{3}$$

Original SOC Cost Function

$$\min_{u \in \mathcal{U}} \mathcal{L}_{SOC} := \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u(X_t^u, t)\|^2 + f(X_t^u, t)\right) dt + g(X_T^u)\right] \tag{4}$$

SOCM Cost Function

$$\min_{u, M} \mathcal{L}_{SOCM}(u, M) := \mathbb{E}\left[\frac{1}{T}\int_0^T \|u(X_t^v, t) - w(t, v, X^v, B, M_t)\|^2 dt \times \alpha(v, X^v, B)\right] \tag{5}$$

Key Point: Same underlying stochastic dynamics, but SOCM uses a least-squares matching approach instead of direct minimization of the original cost.

SOCM Parameters Definition

$$\min_{u,M} \mathcal{L}_{SOCM}(u, M) := \mathbb{E}\left[\frac{1}{T}\int_0^T \|u(X_t^v, t) - w(t, v, X^v, B, M_t)\|^2 dt \times \alpha(v, X^v, B)\right]$$

Where:

$u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the **control** (policy being learned)

v is a **fixed arbitrary control**, X^v is the solution of the SDE with control v

$M : [0,1]^2 \to \mathbb{R}^{d\times d}$ is the **reparameterization matrix**

w is the **matching vector field** (target for u to match)

α is the **importance weight** (for measure correction)

Key Dependencies

w and α depend on f, g, λ, σ

SOCM Parameters Definition

$$\min_{u,M} \mathcal{L}_{SOCM}(u, M) := \mathbb{E}\left[\frac{1}{T}\int_0^T \|u(X_t^v, t) - w(t, v, X^v, B, M_t)\|^2 dt \times \alpha(v, X^v, B)\right]$$

Where:

$u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the **control** (policy being learned)

v is a **fixed arbitrary control**, X^v is the solution of the SDE with control v

$M : [0,1]^2 \to \mathbb{R}^{d\times d}$ is the **reparameterization matrix**

w is the **matching vector field** (target for u to match)

α is the **importance weight** (for measure correction)

Key Dependencies

w and α depend on f, g, λ, σ

Control u: This is what we're trying to learn - the optimal policy that minimizes our cost function.

Arbitrary Control v: Used to generate sample trajectories for training. The choice of v affects the efficiency but not the correctness of SOCM.

Reparameterization Matrix M: Jointly optimized with u to reduce variance in the gradient estimation through path-wise reparameterization.

The SOCM Framework (2/3)

Reparameterization Function $w(t, v, X^v, B, M_t)$

The **target matching function** computed via path-wise reparameterization:

$$
\begin{aligned}
w(t, v, X^v, B, M_t) = \sigma(t)^\top \Bigg(&- \int_t^T M_t(s) \nabla_x f(X_s^v, s)\, ds \; - \; M_t(T) \nabla g(X_T^v) \\
&+ \int_t^T \big(M_t(s) \nabla_x b(X_s^v, s) - \partial_s M_t(s) \big) (\sigma^{-1}(s))^\top v(X_s^v, s)\, ds \\
&+ \sqrt{\lambda} \int_t^T \big(M_t(s) \nabla_x b(X_s^v, s) - \partial_s M_t(s) \big) (\sigma^{-1}(s))^\top dB_s \Bigg)
\end{aligned} \tag{6}
$$

Where $M_t(s)$ is the **reparameterization matrix** (learned jointly with u)

The SOCM Framework (2/3)

Importance Weight $\alpha(v, X^v, B)$

The **importance sampling weight** for measure correction:

$$\alpha(v, X^v, B) = \exp\Bigg(-\frac{1}{\lambda}\int_0^T f(X_t^v, t)\, dt - \frac{1}{\lambda} g(X_T^v) \\ - \frac{1}{\sqrt{\lambda}}\int_0^T \langle v(X_t^v, t), dB_t\rangle - \frac{1}{2\lambda}\int_0^T \|v(X_t^v, t)\|^2 dt \Bigg) \tag{7}$$

Incorporates running costs, terminal costs, and control effort

The SOCM Framework (2/3)

Importance Weight $\alpha(v, X^v, B)$

The **importance sampling weight** for measure correction:

$$\alpha(v, X^v, B) = \exp\Bigg(-\frac{1}{\lambda}\int_0^T f(X_t^v, t)\,dt - \frac{1}{\lambda} g(X_T^v) - \frac{1}{\sqrt{\lambda}}\int_0^T \langle v(X_t^v, t), dB_t\rangle - \frac{1}{2\lambda}\int_0^T \|v(X_t^v, t)\|^2 dt \Bigg) \tag{7}$$

Incorporates running costs, terminal costs, and control effort

Reparameterization Matrix M_t: This matrix enables path-wise reparameterization, a technique to reduce variance in gradient estimation by reparameterizing the stochastic process. It's optimized jointly with the control u to minimize the overall SOCM loss.

Importance Weight α: This exponential weight corrects for the mismatch between the learned control measure and the optimal control measure. However, it can have high variance when costs are large or in high dimensions, which is the main limitation of SOCM.

SOCM Algorithm

Algorithm 2 Stochastic Optimal Control Matching (SOCM)

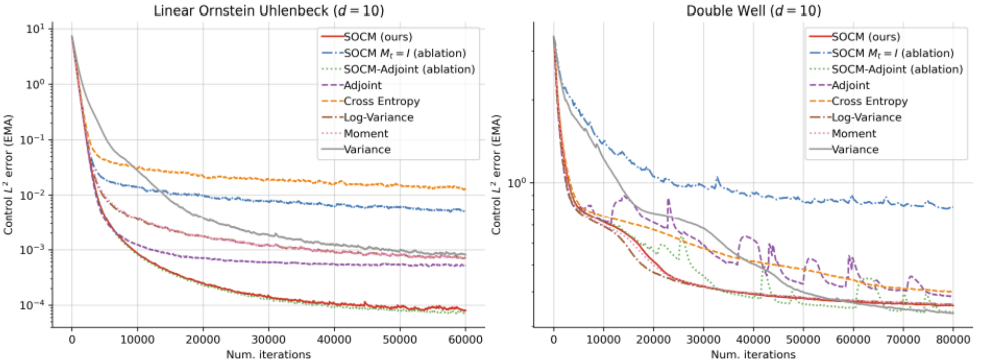
Input: State cost $f(x,t)$, terminal cost $g(x)$, diffusion coeff. $\sigma(t)$, base drift $b(x,t)$, noise level λ, number of iterations N, batch size m, number of time steps K, initial control parameters θ_0, initial matrix parameters ω_0, loss $\mathcal{L}_{\mathrm{SOCM}}$ in (125)

```
for n ∈ {0, ..., N − 1} do
    Simulate m trajectories of the process X^v controlled by v = u_{θ_n}, e.g., using Euler-Maruyama updates
    Detach the m trajectories from the computational graph, so that gradients do not backpropagate
    Using the m trajectories, compute an m-sample Monte-Carlo approximation L̂_SOCM(u_{θ_n}, M_{ω_n}) of the loss L_SOCM(u_{θ_n}, M_{ω_n}) in (125)
    Compute the gradients ∇_{(θ,ω)} L̂_SOCM(u_{θ_n}, M_{ω_n}) of L̂_SOCM(u_{θ_n}, M_{ω_n}) at (θ_n, ω_n)
    Obtain θ_{n+1}, ω_{n+1} with via an Adam update on θ_n, ω_n, resp.
end
```

Output: Learned control u_{θ_N}

Figure: Stochastic Optimal Control Matching (SOCM) Algorithm

Experimental Results (1/2)



Experimental Results (2/2)

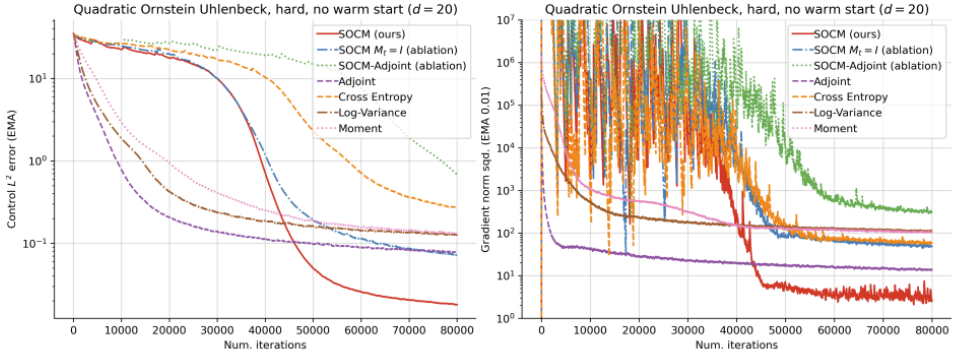


Figure 3 Plots of the L^2 error incurred by the learned control (*top*), and the norm squared of the gradient with respect to the parameters θ of the control (*bottom*), for the QUADRATIC ORNSTEIN UHLENBECK (HARD) setting and for each IDO loss. All the algorithms use a warm-started control (see Appendix D).

Experimental Results (2/2)

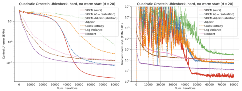


Figure 3 Plots of the L^2 error incurred by the learned control (*top*), and the norm squared of the gradient with respect to the parameters θ of the control (*bottom*), for the Quadratic Ornstein Uhlenbeck (hard) setting and for each IDO loss. All the algorithms use a warm-started control (see Appendix D).

At the end of training, SOCM obtains the lowest L2 error, improving over all existing methods by a factor of around ten. The two SOCM ablations come in second and third by a substantial difference, which underlines the importance of the path-wise reparameterization trick.

JE DOIS COMPRENDRE CE QUE EST UN ORNSTEIN UHLENBECK PROCESS

Conclusion

The main roadblock when we try to apply SOCM to more challenging problems is that the variance of the factor alpha(v, Xv, B) explodes when f and/or g are large, or when the dimension d is high. The control L2 error for the SOCM and cross-entropy losses remains high and fluctuates heavily due to the large variance of alpha The large variance of alpha is due to the mismatch between the probability measures induced by the learned control and the optimal control. Similar problems are encountered in out-of-distribution generalization for reinforcement learning, and some approaches may be carried over from that area (Munos et al., 2016).

References