ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 64 за 2023 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

С.В. Ершов, В.А. Фролов,
А.А. Николаев, А.Г. Волобой

Динамика Ланжевена в стохастической трассировке лучей: выбор фазового пространства и ограничения на вариацию трассы

Рекомендуемая форма библиографической ссылки: Динамика Ланжевена в стохастической трассировке лучей: выбор фазового пространства и ограничения на вариацию трассы / С.В. Ершов [и др.] // Препринты ИПМ им. М.В.Келдыша. 2023. № 64. 15 с.
https://doi.org/10.20948/prepr-2023-64
https://library.keldysh.ru/preprint.asp?id=2023-64

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

С.В. Ершов, В.А. Фролов, А.А. Николаев,
А.Г. Волобой

Динамика Ланжевена в стохастической трассировке лучей: выбор фазового пространства и ограничения на вариацию трассы

Москва — 2023

С.В. Ершов, В.А. Фролов, А.А. Николаев, А.Г. Волобой

Динамика Ланжевена в стохастической трассировке лучей: выбор фазового пространства и ограничения на вариацию трассы

Основная вычислительно емкая задача реалистичной компьютерной графики — расчет глобальной освещенности. В работе проводится исследование скорости сходимости расчета освещения при помощи интегрирования методом Монте-Карло на основе уравнения Ланжевена. В статье представлена вторая часть работы, в которой исследуются выбор фазового пространства, ограничения на возможные вариации трассы и вычисление плотности вероятности предложения перехода. Показано, как эти аспекты влияют на сходимость.

Ключевые слова: глобальная освещенность, стохастическая трассировка лучей, марковские цепи, уравнение Ланжевена.

S.V. Ershov, V.A. Frolov, A.A. Nikolaev, A.G. Voloboy
Langevin dynamics in stochastic ray tracing: phase space selection and limitations for path variation

The main computationally expensive task of realistic computer graphics is the calculation of global illumination. The work investigates the speed of the convergence of lighting simulation using Monte Carlo integration based on the Langevin equation. The paper presents the second part of the work, which analyses the choice of the phase space, restrictions on the possible variations of light path, and the calculation of the probability density of the transition proposal. It is shown how these aspects affect convergence.

Key words: global illumination, stochastic ray tracing, Markov chain, Langevin equation.

1 Введение

Современная реалистичная компьютерная графика базируется на физически корректном моделировании распространения света, в основе которого лежит расчет глобальной освещенности. Расчет освещенности описывается уравнением рендеринга [1], ядром которого является интегрирование значений света, приходящего со всех возможных направлений. Большинство решений этой задачи основываются на стохастической (Монте-Карло) трассировке лучей света. Среди них методы Монте-Карло на основе марковских цепей MCMC (Markov Chain Monte Carlo) [2] становятся все более популярными. В MCMC выборки имеют корреляции между собой. Эта особенность позволяет эффективно использовать информацию об областях функции, имеющих высокую значимость. Один из наиболее распространенных вариантов MCMC — это алгоритм Метрополиса-Гастингса [3]. Основная цель всех алгоритмов MCMC заключается в построении распределения выборок, пропорционального произвольной целевой функции.

В работе [4] говорится о перспективности Монте-Карло методов на основе уравнений Гамильтона и Ланжевена, т.к. эти методы обладают лучшей сходимостью при росте размерности пространства интегрирования. Однако это методы новые, и многие их аспекты, включая вопросы эффективного применения, остаются неисследоваными. Это и послужило мотивацией наших исследований, результаты которых изложены в трех препринтах [5–7].

В первой части работы [5] мы описали общие соображения, касающиеся использования уравнения Ланжевена для генерации случайных трасс лучей. Эффективность этого метода существенно повышается при использовании обобщенного уравнения Ланжевена, включающего матрицу предобработки для учета ограничений на ход луча в сцене. Это позволяет траектории двигаться по разрешенным направлениям с большими смещениями, что дает больше независимых выборок и через это лучшее усреднение в методе Монте-Карло (то есть та же точность оценки яркости пикселя достигается быстрее).

В [5] мы предложили универсальный метод вычисления матрицы предобработки по ограничениям на ход лучей. Все это зависит от выбора фазового пространства, и в данной части мы опишем фазовое пространство, в котором ограничения из-за почти зеркальных ограничений формулируются просто.

2 Выбор фазового пространства

Выбор представления трассы луча, т.е. фазового пространства, существенным образом влияет на все аспекты работы с уравнением Ланжевена для генерирования трасс лучей. В частности, от выбора фазового пространства зависит конкретная форма ограничений на ход лучей в данной сцене, а через это на матрицу предварительной обработки $\hat{T}$ и выбор шага интегрирования по времени [5]. Поэтому правильный выбор фазового пространства чрезвычайно важен, и одним из главных критериев тут является то, сколь хорошо оно для учета *ограничений.*

Применение алгоритма MALA (Metropolis Adjusted Langevin Algorithm) [8–10] оправдано для случая сложного рельефа функции значимости, так как иначе и более простые методы с быстрым шагом работают хорошо. Сложный рельеф возникает, когда оптические свойства объектов сцены близки к зеркальным и (или) свет проходит через маленькие отверстия или, наоборот, отражается от маленьких объектов сцены. Наша модельная сцена [7], созданная для вычислительных экспериментов, имеет обе эти особенности.

Для трассы из $N+1$ сегментов последний из них хранить не надо, так как он однозначно определяется по предыдущим N сегментам (Рис. 1, слева). Для представления трассы направлениями сегментов функция значимости равна произведению значений BDF (в единицах интенсивности) в узлах трассы. Функция BDF (Bidirectional scattering Distribution Function) в общем виде описывает оптические свойства объекта или поверхности сцены. Почти зеркальные BDF приводят к тому, что для значимой трассы направления сегментов не могут сильно отклоняться от направления идеально зеркального отражения предыдущего сегмента $\boldsymbol{s}_i$:

$$\begin{aligned} \boldsymbol{v}_{i+1} &\approx \boldsymbol{s}_i, \\ \boldsymbol{s}_i &= \boldsymbol{v}_i - 2(\boldsymbol{v}_i, \boldsymbol{n}_i)\boldsymbol{n}_i, \end{aligned} \tag{1}$$

где $\boldsymbol{v}_i$ — направление i-го сегмента луча, а $\boldsymbol{n}_i$ — нормаль в *конечной* точке i-го сегмента. Условие $\boldsymbol{v}_{i+1} \approx \boldsymbol{s}_i$ и есть исходная форма *ограничения* из-за BDF сцены.

Поэтому удобно выбрать фазовое пространство, где именно *отклонения* от зеркального направления являются независимыми переменными, т.е. входят в вектор представления трассы луча. Отклонения эти можно описывать по-разному, например, через вектор $\boldsymbol{\omega}_i \perp \boldsymbol{s}_i$, определенный, как показано на правой части Рисунка 1. Для первого сегмента луча нет предыдущего луча и его зеркального отражения, так что первый сегмент надо задавать явно, например, через его направление $\boldsymbol{v}_1$.

По построению, вектор отклонения $\boldsymbol{\omega}_i$ ортогонален к зеркальному направлению $\boldsymbol{s}_i$ и

$$\begin{aligned} \boldsymbol{v}_{i+1} &= \sqrt{1-|\boldsymbol{\omega}_i|^2}\boldsymbol{s}_i + \boldsymbol{\omega}_i, \\ \boldsymbol{\omega}_i &= \boldsymbol{v}_{i+1} - (\boldsymbol{v}_{i+1}, \boldsymbol{s}_i)\boldsymbol{s}_i. \end{aligned} \tag{2}$$

Для представления трассы вектором $\{\boldsymbol{v}_1; \boldsymbol{\omega}_1; \boldsymbol{\omega}_2; \boldsymbol{\omega}_3; ...; \boldsymbol{\omega}_{N-1}\}$ якобиан перехода от набора направлений $\{\boldsymbol{v}_1; \boldsymbol{v}_2; ...; \boldsymbol{v}_N\}$ — единица, так что функция значимости в этом представлении все та же, то есть произведение BDF в единицах

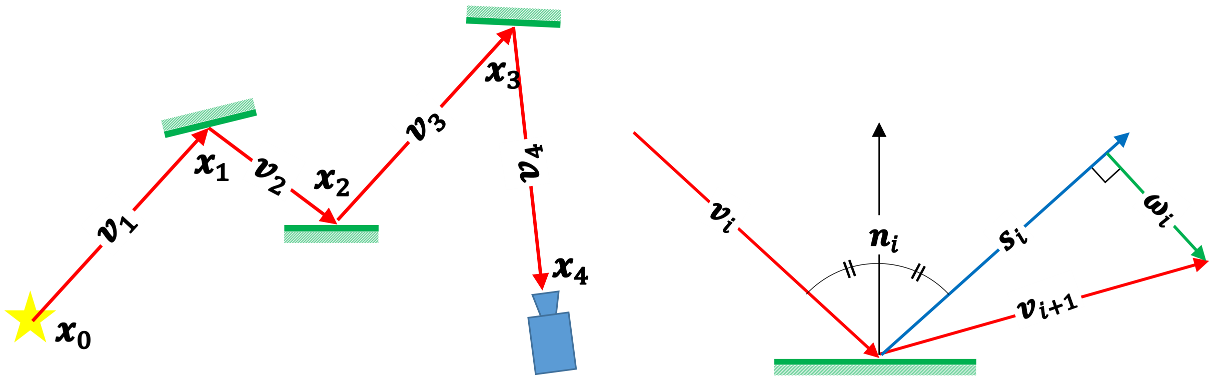

Рис. 1. Левая половина: для трассы из $N+1$ сегментов последний сегмент, заканчивающийся в камере, является функцией предыдущих $\{\boldsymbol{v}_1; \boldsymbol{v}_2; ...; \boldsymbol{v}_N\}$, так как его конец (положение камеры) фиксирован. Показан случай $N = 4$. Правая половина: направление сегмента можно задавать через его *отклонение* от зеркального рассеяния предыдущего сегмента. Красные стрелки — падающий и отраженный сегменты луча, синяя стрелка — зеркальное отражение падающего сегмента, зеленая стрелка — вектор отклонения от зеркального направления, черная — нормаль к поверхности. В алгебраической форме соотношения между векторами показаны в (2).

интенсивности. Входящие в нее $\boldsymbol{v}_i$ вычисляются по $\{\boldsymbol{v}_1; \boldsymbol{\omega}_1; \boldsymbol{\omega}_2; \boldsymbol{\omega}_3; ...; \boldsymbol{\omega}_{i-1}\}$.

Коль скоро $\boldsymbol{\omega}_i$ ортогонален зеркальному направлению $\boldsymbol{s}_i$, то это фактически *двумерный* вектор[1] $\boldsymbol{\Omega}_i = \{\Omega_{i,p}, \Omega_{i,q}\}$, содержащий проекции на какие-то два направления $\boldsymbol{p}'$ и $\boldsymbol{q}'$, перпендикулярные к $\boldsymbol{s}_i$.

Таким образом, представление трассы из N сегментов (точка в фазовом пространстве) теперь есть

$$\boldsymbol{X} = \{\boldsymbol{v}_1; \boldsymbol{\Omega}_1; \boldsymbol{\Omega}_2; \boldsymbol{\Omega}_3; ...; \boldsymbol{\Omega}_{N-1}\}.$$

Вариация $\delta\boldsymbol{X}$, как было объяснено в разделе 2 работы [5], — это вектор в касательной гиперплоскости. Соответствующая вариация *единичного вектора* $\boldsymbol{v}_1$ есть двумерный вектор $\delta\boldsymbol{V}_1 = \{\delta V_p; \delta V_q\}$, содержащий две проекции $\delta\boldsymbol{v}_1$ на направления $\boldsymbol{p}_1$ и $\boldsymbol{q}_1$, перпендикулярные к $\boldsymbol{v}_1$. Таким образом,

$$\delta\boldsymbol{X} = \{\delta\boldsymbol{V}_1; \delta\boldsymbol{\Omega}_1; ...; \delta\boldsymbol{\Omega}_N\}.$$

Для такого представления нельзя вычислять новую точку просто как $\hat{\boldsymbol{X}} = \boldsymbol{X} + \delta\boldsymbol{X}$, ибо тогда первый элемент, $\boldsymbol{v}_1 + \delta\boldsymbol{v}_1$, перестанет быть единичным вектором. Самый простой способ — это принудительно нормировать результат, что есть частный случай проекции на разрешенное многообразие [5]:

[1] Заглавные буквы используются для двухкомпонентного вектора, а строчные буквы остаются за трехкомпонентным вектором в пространстве сцены.

$$\begin{aligned}\hat{\boldsymbol{X}} &= \left\{\frac{\boldsymbol{v}_1+\delta\boldsymbol{v}_1}{|\boldsymbol{v}_1+\delta\boldsymbol{v}_1|};\quad \boldsymbol{\Omega}_1+\delta\boldsymbol{\Omega}_1;\quad \boldsymbol{\Omega}_2+\delta\boldsymbol{\Omega}_2;...;\quad \boldsymbol{\Omega}_{N-1}+\delta\boldsymbol{\Omega}_{N-1}\right\}\\ &= \left\{\frac{\boldsymbol{v}_1+\boldsymbol{p}_1\delta V_p+\boldsymbol{q}_1\delta V_q}{\sqrt{1+\delta V_p^2-\delta V_q^2}};\quad \boldsymbol{\Omega}_1+\delta\boldsymbol{\Omega}_1;\right.\\ &\qquad \left.\boldsymbol{\Omega}_2+\delta\boldsymbol{\Omega}_2;...;\quad \boldsymbol{\Omega}_{N-1}+\delta\boldsymbol{\Omega}_{N-1}\right\}. \end{aligned}\tag{3}$$

Возможны и другие способы, например, взять новое направление как $\hat{\boldsymbol{v}}_1 = \sqrt{1-|\delta\boldsymbol{v}_1|^2}\boldsymbol{v}_1+\delta\boldsymbol{v}_1$.

Функция значимости в этом пространстве (как и в пространстве, где трасса описана направлениями всех сегментов) оказывается просто произведением значений BDF (в единицах интенсивности) в узлах трассы.

ЗАМЕЧАНИЕ. Очевидно, $\boldsymbol{\omega}_i$, и потому $\boldsymbol{v}_{i+1}$ зависят и от двух чисел $\{\Omega_{i,p},\Omega_{i,q}\}$, и от координатной системы $\boldsymbol{p}'_i,\boldsymbol{q}'_i$, в свою очередь определяющуюся i-м (=предшествующим) сегментом. А он, в свою очередь, зависит и от $\{\Omega_{i-1,p},\Omega_{i-1,q}\}$, и от $i-1$-го сегмента, ..., и так далее.

$$\boldsymbol{\omega}_i = \Omega_{i,p}\boldsymbol{p}'_i(\boldsymbol{v}_1,\boldsymbol{\Omega}_1,...,\boldsymbol{\Omega}_{i-1})+\Omega_{i,q}\boldsymbol{q}'_i(\boldsymbol{v}_1,\boldsymbol{\Omega}_1,...,\boldsymbol{\Omega}_{i-1}).\tag{4}$$

В итоге, хотя бы компонента $\boldsymbol{\Omega}_i$ и *была фиксирована*, соответствующий вектор $\boldsymbol{\omega}_i$ может меняться. Матрица Якоби преобразования $\{\boldsymbol{v}_1;\boldsymbol{\Omega}_1;\boldsymbol{\Omega}_2;\boldsymbol{\Omega}_3;...;\boldsymbol{\Omega}_{N-1}\}\mapsto\{\boldsymbol{v}_1;\boldsymbol{\omega}_1;\boldsymbol{\omega}_2;\boldsymbol{\omega}_3;...;\boldsymbol{\omega}_{N-1}\}$ — нижняя треугольная, поэтому определитель ее есть произведение диагональных элементов и оказывается равен 1. Так что и в таком представлении функция значимости остается прежним произведением значений BDF.

3 Плотности вероятности предложения прямого и обратного перехода

Мы выбрали фазовое пространство, в котором отклонения от зеркального направления являются независимыми переменными, т.е. входят в вектор представления трассы луча. Для реализации MALA необходимо знать плотность вероятности предложения перехода в этом фазовом пространстве, т.е. плотность вероятности, с какой распределена $\hat{\boldsymbol{X}}$ при фиксированном $\boldsymbol{X}$, но случайном $\boldsymbol{\xi}$). Отображение это задается двумя формулами: формула (4) из работы [5], по какой находим $\delta\boldsymbol{X}=\{\delta\boldsymbol{V}_1;\delta\boldsymbol{\Omega}_1;...;\delta\boldsymbol{\Omega}_N\}$, и формула (3), по которой уже находим $\hat{\boldsymbol{X}}$.

Итак, условная плотность распределения $p(\hat{\boldsymbol{X}}|\boldsymbol{X})$ относится к случаю, когда $\boldsymbol{X}$ фиксировано, и случайность $\hat{\boldsymbol{X}}$ обусловлена случайным вектором $\boldsymbol{\xi}$.

А коль скоро случайная величина $\hat{\boldsymbol{X}}$ есть функция случайной величины $\boldsymbol{\xi}$, распределенной по Гауссу с плотностью G, то и плотность распределения $\hat{\boldsymbol{X}}$ вычисляется как

$$p(\hat{\boldsymbol{X}}|\boldsymbol{X}) = \frac{1}{\left|\det \frac{\partial \hat{\boldsymbol{X}}}{\partial \boldsymbol{\xi}}\right|} G(\boldsymbol{\xi}), \tag{5}$$

где $\boldsymbol{\xi}$ должно быть вычислено по $\hat{\boldsymbol{X}}$ и $\boldsymbol{X}$. А именно, сперва из (3) находим $\delta \boldsymbol{X} = \{\delta \boldsymbol{V}_1; \delta \boldsymbol{\Omega}_1; ...; \delta \boldsymbol{\Omega}_N\}$ по формулам

$$\delta \boldsymbol{v}_1 = \frac{\hat{\boldsymbol{v}}_1}{(\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1)} - \boldsymbol{v}_1, \tag{6}$$

$$\delta \boldsymbol{V}_1 = \{(\boldsymbol{p}_1, \delta \boldsymbol{v}_1); (\boldsymbol{q}_1, \delta \boldsymbol{v}_1)\}, \tag{7}$$

$$\delta \boldsymbol{\Omega}_i = \hat{\boldsymbol{\Omega}}_i - \boldsymbol{\Omega}_i \tag{8}$$

и затем вычисляем уже $\boldsymbol{\xi}$ как:

$$\boldsymbol{\xi} = \frac{1}{\sqrt{2\delta t}} \hat{T}^{-1}(\boldsymbol{X}) \left(\delta \boldsymbol{X} - \boldsymbol{a}(\boldsymbol{X}) \delta t\right). \tag{9}$$

Определитель матрицы Якоби $\left|\det \frac{\partial \hat{\boldsymbol{X}}}{\partial \boldsymbol{\xi}}\right|$ вычисляется как производная комбинации преобразований $\left|\det \frac{\partial \hat{\boldsymbol{X}}}{\partial \boldsymbol{\xi}}\right| = \left|\det \frac{\partial (\delta \boldsymbol{X})}{\partial \boldsymbol{\xi}} \det \frac{\partial \hat{\boldsymbol{X}}}{\partial (\delta \boldsymbol{X})}\right|$. Первый член находится из формулы (4) из работы [5]: $\det \frac{\partial \hat{\boldsymbol{X}}}{\partial (\delta \boldsymbol{X})} = (\sqrt{2\delta t})^N \left|\det \hat{T}\right|$, а второй — из преобразования (3):

$$\det \frac{\partial (\hat{\boldsymbol{v}}_1; \hat{\boldsymbol{\Omega}}_1; ...; \delta \hat{\boldsymbol{\Omega}}_N)}{\partial (\delta \boldsymbol{V}_1; \delta \boldsymbol{\Omega}_1; ...; \delta \boldsymbol{\Omega}_N)} = \det \begin{pmatrix} \frac{\partial (\hat{\boldsymbol{v}}_1)}{\partial (\delta \boldsymbol{V}_1)} & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \det \frac{\partial (\hat{\boldsymbol{v}}_1)}{\partial (\delta \boldsymbol{V}_1)},$$

где $\hat{\boldsymbol{v}}_1 = \frac{\boldsymbol{v} + \boldsymbol{p}\delta V_p + \boldsymbol{q}\delta V_q}{\sqrt{1 + \delta V_p^2 - \delta V_q^2}}$. Матрица Якоби $\frac{\partial \frac{\boldsymbol{v} + \boldsymbol{p}\delta V_p + \boldsymbol{q}\delta V_q}{\sqrt{1 + \delta V_p^2 - \delta V_q^2}}}{\partial (\delta V_p; \delta V_q)}$ есть, очевидно, $(\hat{\boldsymbol{v}}, \boldsymbol{v}) \begin{pmatrix} (\boldsymbol{p}_1, \hat{\boldsymbol{p}}_1) & (\boldsymbol{q}_1, \hat{\boldsymbol{p}}_1) \\ (\boldsymbol{p}, \hat{\boldsymbol{q}}) & (\boldsymbol{q}_1, \hat{\boldsymbol{q}}_1) \end{pmatrix}$, где $\hat{\boldsymbol{p}}_1$ и $\hat{\boldsymbol{q}}_1$ — аналоги $\boldsymbol{p}_1$ и $\boldsymbol{q}_1$, то есть направления, ортогональные к $\hat{\boldsymbol{v}}_1$. Таким образом,

$$\left|\det \frac{\partial \hat{\boldsymbol{X}}}{\partial \boldsymbol{\xi}}\right| = (\sqrt{2\delta t})^N \left|\det \hat{T}\right| (\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1)^2 \left|(\boldsymbol{p}_1, \hat{\boldsymbol{p}}_1)(\boldsymbol{q}_1, \hat{\boldsymbol{q}}_1) - (\boldsymbol{q}_1, \hat{\boldsymbol{p}}_1)(\boldsymbol{p}_1, \hat{\boldsymbol{q}}_1)\right|$$

и

$$p(\hat{\boldsymbol{X}}|\boldsymbol{X}) = \frac{G(\boldsymbol{\xi})}{(\sqrt{2\delta t})^N \left|\det \hat{T}(\boldsymbol{X})\right| (\boldsymbol{v}_1, \hat{\boldsymbol{v}}_1)^2 |(\hat{\boldsymbol{p}}_1, \boldsymbol{p}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{q}_1) - (\hat{\boldsymbol{p}}_1, \boldsymbol{q}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{p}_1)|}, \tag{10}$$

где $\boldsymbol{\xi}$ дается формулой (9).

Плотность вероятности перехода в *обратном направлении* есть, конечно же, та же самая функция, но с переставленными аргументами:

$$p(\hat{\boldsymbol{X}}|\boldsymbol{X}) = \frac{G(\boldsymbol{\eta})}{(\sqrt{2\delta t'})^N \left|\det \hat{T}(\boldsymbol{X})\right| (\boldsymbol{v}_1, \hat{\boldsymbol{v}}_1)^2 |(\hat{\boldsymbol{p}}_1, \boldsymbol{p}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{q}_1) - (\hat{\boldsymbol{p}}_1, \boldsymbol{q}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{p}_1)|},$$

где $\delta t'$ находится для *исходной точки шага* в $\hat{\boldsymbol{X}}$, а

$$\boldsymbol{\eta} = \frac{1}{\sqrt{2\delta t'}} \hat{T}^{-1}(\hat{\boldsymbol{X}}) \left(\delta \boldsymbol{X}' - \boldsymbol{a}(\hat{\boldsymbol{X}})\delta t\right)$$

и, наконец, $\delta \boldsymbol{X}' = \{\delta \boldsymbol{V}'_1; \delta \boldsymbol{\Omega}'_1; ...; \delta \boldsymbol{\Omega}'_N\}$ есть смещение точки в касательной гиперплоскости при воображаемом обратном переходе из $\hat{\boldsymbol{X}}$ в $\boldsymbol{X}$, и потому вычисляется точно как и для прямого перехода, просто переставлением $\boldsymbol{X} \leftrightarrow \hat{\boldsymbol{X}}$ в формулах (6)–(8).

Заметим, что член $(\boldsymbol{v}_1, \hat{\boldsymbol{v}}_1)^2 |(\hat{\boldsymbol{p}}_1, \boldsymbol{p}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{q}_1) - (\hat{\boldsymbol{p}}_1, \boldsymbol{q}_1)(\hat{\boldsymbol{q}}_1, \boldsymbol{p}_1)|$ в знаменателе выражений для $p(\hat{\boldsymbol{X}}|\boldsymbol{X})$ и $p(\boldsymbol{X}|\hat{\boldsymbol{X}})$ — *один и тот же* (из-за его симметрии относительно перестановки $\boldsymbol{X} \leftrightarrow \hat{\boldsymbol{X}}$), так что он сокращается при вычислении вероятности принятия в правиле Гастингса (формула (6) в работе [5]).

4 Ограничения на вариацию трассы

4.1 Пространственные ограничения в мировом пространстве путей

Часто рельеф функции значимости в фазовом пространстве содержит узкие «овраги», и траектория должна «пробираться» внутри них (как в методах спуска[2]). Если овраг имеет гладкие склоны, то градиент потенциала тянет траекторию к средней линии оврага. Для узкого оврага градиент велик, так что необходимо сильно ограничивать δt, чтобы не выскочить из оврага. Но при малом δt движение вдоль оврага слишком медленное, так что нужно очень большое число шагов, чтобы пройти его весь (да не по одному разу, чтобы набрать статистику).

Ситуация сходна с методами градиентного спуска, и, как и там, положение значительно улучшается при применении матрицы локального преобразования переменных, именуемой “preconditioning matrix” в алгоритмах гибридного

[2]Между этими методами действительно есть содержательная аналогия [11].

Монте-Карло. Эта матрица делает малой проекцию $\delta\boldsymbol{X}$ на направление поперек оврага и большой — на направление вдоль него. Можно сказать, что она адаптирует предложение смещения $\delta\boldsymbol{X}$ к рельефу.

В трассировке лучей такие узкие овраги с плавными склонами возникают в основном из-за наличия глянцевых поверхностей, чей BDF быстро исчезает при отклонении луча от направления зеркального отражения или преломления. Другой случай — овраги с вертикальными склонами, т.е. узкие «потенциальные ямы», или хотя бы значительные конечные *разрывы* потенциала. Они могут возникать, например, в случае, когда луч рассеивается маленьким объектом сцены или проходит сквозь маленькое отверстие (рис. 2).

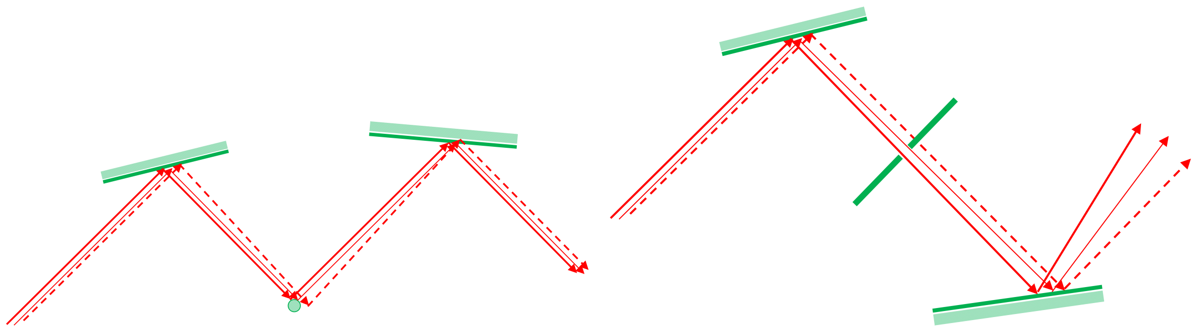

Рис. 2. Два случая, когда возникает узкая потенциальная яма. Слева: луч отражается от маленького объекта сцены. Справа: луч проходит сквозь маленькое отверстие. Даже небольшая вариация трассы может привести к тому, что она станет невозможной. Жирные красные стрелки показывают сегменты исходной трассы луча, тонкие стрелки показывают возможные мутации, т.е. слегка смещенную трассу, а пунктирные стрелки показывают невозможные мутации, т.е. такая трасса невозможна.

Случай прохождения сквозь маленькое отверстие наиболее интуитивно понятен. Пусть узлы трассы луча перед и за отверстием — это $\boldsymbol{x}_k$ и $\boldsymbol{x}_{k+1}$ (очевидно, что это начало и конец одного сегмента), а отверстие лежит в плоскости $(\boldsymbol{x},\boldsymbol{n}) = c$. Тогда трасса пересекает эту плоскость в точке

$$\boldsymbol{x} = \boldsymbol{x}_k + \frac{c - (\boldsymbol{x}_k,\boldsymbol{n})}{((\boldsymbol{x}_{k+1},\boldsymbol{n}) - (\boldsymbol{x}_k,\boldsymbol{n}))}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$$

и бесконечно малое смещение узлов $\boldsymbol{x}_k$ и $\boldsymbol{x}_{k+1}$ приводит к следующей вариации точки пересечения плоскости:

$$\begin{aligned} d\boldsymbol{x} \;\; = \;\; & -\frac{c - (\boldsymbol{x}_{k+1},\boldsymbol{n})}{((\boldsymbol{x}_{k+1},\boldsymbol{n}) - (\boldsymbol{x}_k,\boldsymbol{n}))} d\boldsymbol{x}_k + \frac{c - (\boldsymbol{x}_{k+1},\boldsymbol{n})}{((\boldsymbol{x}_{k+1},\boldsymbol{n}) - (\boldsymbol{x}_k,\boldsymbol{n}))^2}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)(d\boldsymbol{x}_k,\boldsymbol{n}) \\ & + \frac{c - (\boldsymbol{x}_k,\boldsymbol{n})}{((\boldsymbol{x}_{k+1},\boldsymbol{n}) - (\boldsymbol{x}_k,\boldsymbol{n}))} d\boldsymbol{x}_{k+1} - \frac{c - (\boldsymbol{x}_k,\boldsymbol{n})}{((\boldsymbol{x}_{k+1},\boldsymbol{n}) - (\boldsymbol{x}_k,\boldsymbol{n}))^2}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)(d\boldsymbol{x}_{k+1},\boldsymbol{n}), \end{aligned}$$

то есть

$$d\boldsymbol{x} = \hat{A}(\boldsymbol{x}_k;\boldsymbol{x}_{k+1})d\boldsymbol{x}_k + \hat{A}(\boldsymbol{x}_{k+1};\boldsymbol{x}_k)d\boldsymbol{x}_{k+1}.$$

Очевидно, что в этом случае в 4-мерном касательном подпространстве $(d\boldsymbol{x}_k; d\boldsymbol{x}_{k+1})$ есть направления, смещение по которым не сдвигает $\boldsymbol{x}$ вовсе, так что даже большая вариация трассы в этом направлении оставляет ее в разрешенной области фазового пространства. Есть направления, смещение по которым сдвигает точку прохождения сквозь отверстие вдоль длинной стороны отверстия, так что в этом направлении трассу сместить можно на заметный (но все же небольшой) шаг, пока она не выйдет из разрешенной области. И, наконец, есть направления, сдвигающие точку прохождения сквозь отверстие вдоль его короткой стороны — тут уже допустимый шаг самый маленький.

Пусть $\{\boldsymbol{e}_i\}$ — набор ортонормированных векторов, адаптированный к ориентации отверстия, и l_i — размер отверстия вдоль $\boldsymbol{e}_i$. Тогда ограничение на смещение точки пересечения лучом плоскости отверстия

$$(d\boldsymbol{x}, \boldsymbol{e}_i)^2 \leq l_i^2, \qquad i = 1, ..., N,$$

то есть

$$(\hat{A}(\boldsymbol{x}_k; \boldsymbol{x}_{k+1}) d\boldsymbol{x}_k + \hat{A}(\boldsymbol{x}_{k+1}; \boldsymbol{x}_k) d\boldsymbol{x}_{k+1}, \boldsymbol{e}_i)^2 \leq l_i^2, \qquad i = 1, ..., N. \tag{11}$$

Случай, когда узел трассы луча попадает на маленький объект сцены, рассматривается аналогично и даже проще. Допустим, что на этот объект попадает узел трассы луча $\boldsymbol{x}_k$. Тогда разрешены те мутации, при которых $\boldsymbol{x}_k + d\boldsymbol{x}_k$ все еще находится на поверхности этого тела. Приближенно это означает, что смещение $d\boldsymbol{x}_k$ (а оно лежит в плоскости, касательной к поверхности объекта в точке $\boldsymbol{x}_k$) должно быть ограничено, скорее всего, по-разному в разных направлениях. Общий вид его таков:

$$(d\boldsymbol{x}_k, \hat{B}(\boldsymbol{x}_k) d\boldsymbol{x}_k) \leq 1. \tag{12}$$

4.2 Ограничения для выбранного фазового пространства

Ниже дается подробное описание того, как матрица преобразования $\hat{T}$ вычисляется по ограничениям для фазового пространства, описанного в Главе 2.

4.2.1 Ограничения из-за оптических свойст объектов

Отклонения от зеркального направления, которые и ограничиваются для BDF, близким к зеркальному отражению, являются независимыми переменными, входящими в представление трассы луча для этого пространства. Поэтому ограничения на них следующие:

$$|\boldsymbol{\Omega}_i| \leq \max \vartheta_i, \tag{13}$$

где $\max\vartheta_i$ — ширина лепестка BDF, за пределами которого BDF практически обращается в 0. Заметим, что само фазовое пространство ограничений на эту переменную не налагает и любой вектор $\boldsymbol{\Omega}_i$ допустим. Другое дело, что значения BDF для него будут черезвычайно малы.

Мы ограничим вариацию исходной трассы так, что

$$(\delta\boldsymbol{\Omega}_i, \delta\boldsymbol{\Omega}_i) \leq (\max\vartheta_i)^2, \tag{14}$$

тогда после вариации в худшем случае будет $|\boldsymbol{\Omega}_i| \leq 2\max\vartheta_i$. Множитель 2 не очень существенен потому, что хотя теперь *некоторые* вариации трассы и приведут к тому, что BDF в каком-то узле *почти* 0, но все равно значительная доля случайных мутаций приведет к разрешенным трассам с ненулевым значениям BDF.

Однако значение $\boldsymbol{\Omega}_N$ в завершающем сегменте, заканчивающимся в источнике, не является независимой переменной и даже не входит в вектор состояния. Аналогично и $d\boldsymbol{\Omega}_N$ не входит в $d\boldsymbol{X}$, но вычисляется по нему, то есть из $(d\boldsymbol{v}_1; d\boldsymbol{\Omega}_1; d\boldsymbol{\Omega}_2; d\boldsymbol{\Omega}_3; ...; d\boldsymbol{\Omega}_{N-1})$. Общая форма этой зависимости, очевидно, такова:

$$d\boldsymbol{\Omega}_N = \hat{\mathcal{B}} d\boldsymbol{X} = \hat{B}_0 d\boldsymbol{V}_1 + \hat{B}_1 d\boldsymbol{\Omega}_1 + \hat{B}_2 d\boldsymbol{\Omega}_2 + \cdots + \hat{B}_{N-1} d\boldsymbol{\Omega}_{N-1}. \tag{15}$$

Это соотношение приближенно выполнено и для конечных вариаций, поэтому ограничение $|\delta\boldsymbol{\omega}_N| \leq \max\vartheta_N$ приводит к:

$$|\hat{B}_0 \delta\boldsymbol{V}_1 + \hat{B}_1 \delta\boldsymbol{\Omega}_1 + \hat{B}_2 \delta\boldsymbol{\Omega}_2 + \cdots + \hat{B}_{N-1} \delta\boldsymbol{\Omega}_{N-1}| \leq \max\vartheta_N,$$

что есть квадратичная форма

$$(\delta\boldsymbol{X}, \hat{Q}_1 \delta\boldsymbol{X}) \leq (\max\vartheta_N)^2. \tag{16}$$

Может показаться, что ограничения, записываемые в матрицу предварительной обработки, функционально делают то же самое, что и такие методы улучшения сходимости Монте-Карло, как выборка по значимости [12] и Path Guiding [13]. Мы считаем, что это не совсем так, поскольку упомянутые методы могут быть использованы вместе с динамикой Ланжевена, а не вместо ее (то есть они ортогональны друг другу). При этом сама динамика Ланжевена может работать, например, в первичном пространстве путей, как и было реализовано у Fujun Luang в [8].

4.2.2 Пространственные ограничения

Наиболее распространенные пространственные ограничения таковы:

1. Вариации узлов трассы луча должны быть ограничены так, чтобы нормаль в этих точках менялась не слишком сильно и чтобы точка не уходила с поверхности. Разрешенный размер смещения (узла трассы) обычно разный в разных направлениях.

2. Вариация точки пересечения с плоскостью отверстия должна оставаться внутри этого отверстия

Первый тип ограничения применяется к каждому узлу отдельно, и может быть записан в виде квадратичной формы

$$(\delta \boldsymbol{x}_k, \hat{D}(\boldsymbol{x}_k)\delta \boldsymbol{x}_k) \le d_k^2. \tag{17}$$

Для нашего фазового пространства из Главы 2 координаты узлов трассы не являются независимыми переменными, их надо вычислять по вектору состояния $\boldsymbol{X}$. Поэтому

$$d\boldsymbol{x}_k = \hat{\mathcal{V}}_k d\boldsymbol{X}, \tag{18}$$

причем матрица $\hat{\mathcal{V}}_k$ в общем случае не квадратная. Это соотношение приближенно выполнено и для конечных вариаций, и, подставляя его в (17), имеем

$$\begin{aligned} \hat{Q}_k &\equiv \hat{\mathcal{V}}_k^* \hat{D}(\boldsymbol{x}_k)\hat{\mathcal{V}}_k, \\ (\delta \boldsymbol{X}, \hat{Q}_k \delta \boldsymbol{X}) &\le d_k^2. \end{aligned} \tag{19}$$

Второй тип ограничения является, скорее, даже *связью* между вариациями двух узлов: их надо двигать так, чтобы точка прохождения сквозь отверстие либо стояла на месте, либо смещалась в пределах его. Тогда получается выражение вида (11). Оно верно, если сами смещения узлов трассы малы, но это так и есть ввиду ограничений первого типа. Выражая $d\boldsymbol{x}_k$ из $d\boldsymbol{X}$ по формуле (18) и применяя результат к конечным вариациям, имеем

$$(\delta \boldsymbol{X}, \boldsymbol{w}_{i,k})^2 \le l_i^2, \qquad i = 1, ..., N,$$

где

$$\begin{aligned} \hat{W}_k &\equiv \hat{A}(\boldsymbol{x}_k; \boldsymbol{x}_{k+1})\hat{\mathcal{V}}_k + \hat{A}(\boldsymbol{x}_{k+1}; \boldsymbol{x}_k)\hat{\mathcal{V}}_{k+1}, \\ \boldsymbol{w}_{i,k} &\equiv \hat{W}_k^* \boldsymbol{e}_i^{(k)}. \end{aligned}$$

Квадратичная форма $(\delta \boldsymbol{X}, \boldsymbol{w}_{i,k})^2$ может быть записана и в более привычном матричном виде:

$$\begin{aligned} (\delta \boldsymbol{X}, \hat{Q}_{i,k} \delta \boldsymbol{X}) &\le l_i^2, \qquad i = 1, ..., N, \\ \hat{Q}_{i,k} &\equiv |\boldsymbol{w}_{i,k}\rangle\langle \boldsymbol{w}_{i,k}|. \end{aligned} \tag{20}$$

4.2.3 Объединяем ограничения

В общем случае набор ограничений включает (14), (16), (19) и (20). Все могут быть записаны в виде: квадратичная форма не превосходит заданного предела

$$(\delta \boldsymbol{X}, \hat{Q}_k(\boldsymbol{X})\delta \boldsymbol{X}) \leq q_k^2,$$

где $\hat{Q}_k(\boldsymbol{X})$ — симметричная неотрицательная матрица.

Разумеется, для какой-то сцены часть из этих ограничений может отсутствовать, скажем, если все BDF достаточно плавные или если все отверстия велики и пр. Формально можно сохранить все ограничения, но для фактически отсутствующих задать очень большое (или просто бесконечное) q_k.

Разрешенные вариации трассы луча должны удовлетворять *всем* ограничениям. Простейший способ объединить их в одно — это записать все соответствующие неравенства в форме $(\delta \boldsymbol{X}, q_k^{-2}\hat{Q}_k(\boldsymbol{X})\delta \boldsymbol{X}) \leq 1$ и просуммировать, что приводит к :

$$\begin{aligned} (\delta \boldsymbol{X}, \hat{Q}(\boldsymbol{X})\delta \boldsymbol{X}) &\leq 1, \\ \hat{Q}(\boldsymbol{X}) &\equiv \sum_{k=1}^{M} q_k^{-2}\hat{Q}_k(\boldsymbol{X}). \end{aligned} \tag{21}$$

Это, конечно, излишне сильное ограничение, т.е. в реальности точка может остаться в разрешенной области и при смещении на $\delta \boldsymbol{X}$, заметно большее, чем дозволено (21), что очевидно для случая идентичных ограничений. В качестве некоего компромисса можно было бы сформировать итоговую чуть иначе, например,

$$\hat{Q}(\boldsymbol{X}) = \frac{1}{M}\sum_k q_k^{-2}\hat{Q}_k(\boldsymbol{X}),$$

где M — количество исходных, индивидуальных ограничений. Если бы они все были одинаковы, такая матрица обеспечивала бы идеально точное ограничение, в отличие от завышенного требования из (21).

В случае же различающихся исходных ограничений такая матрица может приводить к тому, что предложенное $\delta \boldsymbol{X}$ окажется в $\sqrt{M}$ раз больше, чем дозволено. Такое предложение перехода будет, конечно же, отвергнуто. Однако многие предложения все же будут приняты и дадут большее смещение траектории. Заранее неочевидно, что выгоднее: мелкие шаги, все из которых приняты, или крупные, но часть отвергнута. Поэтому можно предложить формировать итоговую матрицу как

$$\hat{Q}(\boldsymbol{X}) = \frac{1}{\sqrt{M}}\sum_k q_k^{-2}\hat{Q}_k(\boldsymbol{X}).$$

Однако так или иначе мы все равно приходим к неравенству на квадратичную форму наподобие (21), разве что с чуть иначе вычисленной итоговой матрицей $\hat{Q}(\boldsymbol{X})$.

5 Заключение

Стохастическая трассировка лучей с использованием динамики Ланжевена выглядит весьма перспективным методом для расчета глобальной освещенности виртуальной сцены. Однако на сегодняшний день этот метод остается малоисследованным, в то время как его анализ может привести к значительному ускорению вычислений без потери качества получаемого результата.

В данной работе мы продолжили анализ наиболее вычислительно емких операций алгоритма MALA, начатый в [5]. В представленной, второй, части работы основное внимание было уделено выбору представления трассы луча, т.е. фазового пространства, вычислению плотности вероятности предложения перехода в этом фазовом пространстве и анализу ограничений на возможные вариации трассы, необходимых для вычисления матрицы предварительной обработки. Показано, как эти аспекты влияют на сходимость.

Дальнейшие работы [7] будут направлены на проведение вычислительных экспериментов, подтверждающих результаты наших исследований, и анализ их результатов.

Список литературы

[1] Kajiya J.T. The rendering equation // Proceedings of the 13th annual conference on Computer graphics and interactive techniques (ACM SIGGRAPH 86). 1986. 20(4), pp. 143-150.

[2] Sik M., Krivanek J. Survey of Markov Chain Monte Carlo Methods in Light Transport Simulation // IEEE Transactions on Visualization and Computer Graphics. 2018. T. 26. № 4. pp. 1821-1840.

[3] Brooks S., Gelman A., Jones G., Meng X. Handbook of Markov Chain Monte Carlo. CRC press. 2011.

[4] Фролов В.А, Волобой А.Г., Ершов С.В., Галактионов В.А. Современное состояние методов расчета глобальной освещенности в задачах реалистичной компьютерной графики // Труды Института системного программирования РАН. 2021. Т. 33. № 2. С. 7-48. http://dx.doi.org/10.15514/ISPRAS-2021-33(2)-1

[5] Ершов С.В., Фролов В.А., Николаев А.А., Волобой А.Г. Динамика Ланжевена в стохастической трассировке лучей: вычисление матрицы предобработки по ограничениям и выбор шага по времени // Препринты ИПМ им. М.В.Келдыша. 2023. № 63. https://doi.org/10.20948/prepr-2023-63

[6] Ершов С.В., Фролов В.А., Николаев А.А., Волобой А.Г. Динамика Ланжевена в стохастической трассировке лучей: выбор фазового пространства и ограничения на вариацию трассы // Препринты ИПМ им. М.В.Келдыша. 2023. № 64. https://doi.org/10.20948/prepr-2023-64

[7] Ершов С.В., Фролов В.А., Николаев А.А., Волобой А.Г. Динамика Ланжевена в стохастической трассировке лучей: вычислительные эксперименты // Препринты ИПМ им. М.В.Келдыша. 2023. № 65. https://doi.org/10.20948/prepr-2023-65

[8] Luan F., Zhao S., Bala K., Gkioulekas I. Langevin Monte Carlo rendering with gradient-based adaptation // ACM Transactions on Graphics (TOG). 2020. T. 39. № 4, Article 140, 16 pages.

[9] Luan F. Forward and inverse rendering with gradient based optimizations. — Ph.D. Thesis. Cornell University. 2021.

[10] Xifara T., Sherlock C., Livingstone S., Byrne S., Girolami M. Langevin diffusions and the Metropolis-adjusted Langevin algorithm // Statistics & Probability Letters. 2014. T. 91. C. 14-19. arXiv preprint arXiv:1309.2983.

[11] Kingma D.P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

[12] Veach E. Robust Monte Carlo methods for light transport simulation. — Ph.D. Thesis. Stanford University. 1998.

[13] Muller T., Gross M., Novak J. Practical Path Guiding for Efficient Light-Transport Simulation // Computer Graphics Forum (Eurographics 2017). 2017. T. 36. № 4. pp. 91-100.