УДК 550.34
DOI 10.33764/2618-981X-2024-2-2-61-67

Е. А. Гондюл[1], В. В. Лисица[1], К. Г. Гадыльшин[1], Д. М. Вишневский[1]*

Формирование репрезентативной выборки для обучения нейронной сети, подавляющей численную дисперсию в псевдо-трёхмерном пространстве

[1] Институт нефтегазовой геологии и геофизики им. А.А. Трофимука Сибирского отделения Российской академии наук (ИНГГ СО РАН), г. Новосибирск, Российская Федерация
* e-mail: gondyulea@ipgg.sbras.ru

Аннотация. В работе обсуждается использование NDM-net (Numerical Dispersion Mitiagtion neural network) для трёхмерного сейсмического моделирования и формирования обучающей выборки. Изначально, NDM-net была разработана для подавления численной дисперии в сейсмических данных, т.е. в результатах моделирования динамической теории упругости. Предварительно для формирования обучающей выборки рассчитываются сейсмограммы на крупной сетке с численной дисперсией и определенное количество сейсмограмм на мелкой сетке. В статье рассматриваются три подхода к построению репрезентативной выборки для ускорения процесса обучения в псевдо-трёхмерном случае. Кроме того, в статье обсуждается комбинация показателей, основанных на статистическом анализе. Показано, что предложенный метод подавляет численную дисперсию в несколько раз при существенном ускорении моделирования.

Ключевые слова: сейсмическое моделирование, численная дисперсия, нейронные сети

E. A. Gondyul[1], V. V. Lisitsa[1], K. G. Gadylshin[1], D. M. Vishnevsky[1]*

Creation of the representative sample for training numerical dispersion mitigation neural network in pseudo-3D space

[1] Trofimuk Institute of Petroleum Geology and Geophysics of Siberian Branch Russian Academy of Sciences (IPGG SB RAS), Novosibirsk, Russian Federation
* e-mail: gondyulea@ipgg.sbras.ru

Abstract. The paper discusses the application of NDM-net (Numerical Dispersion Mitigation Neural Network) for three-dimensional seismic simulation and the generation of a training dataset. Initially, NDM-net was designed to mitigate numerical dispersion in seismic data, that is, the results of dynamic elastic wave simulations. To generate a training sample, a large number of seismograms with numerical dispersion are computed, as well as a smaller number of seismograms without numerical dispersion on a fine grid. The paper proposes three approaches to creating a representative dataset to accelerate the learning process for the pseudo-3D case. Additionally, the paper considers a combination of metrics based on statistical analysis. The proposed method is shown to significantly reduce numerical variance while accelerating the simulation process.

Keywords: seismic modelling, numerical dispersion, neural network

Введение

Численная дисперсия является одной из проблем, возникающих при численном моделировании сейсмических полей с использованием грубой сетки для ускорения процесса моделирования. Существуют различные способы подавле-

ния численной ошибки, включая классические методы такие, как увеличение порядка аппроксимации; схемы, подавляющие численную дисперсию [1]; разрывный метод Галёркина [2] и т.д. Также используются неклассические методы, которые применяются после моделирования. Однако, все эти методы могут быть трудоёмкими, уменьшение числа Куранта, ужесточение условий стабильности, увеличение числа операций с плавающей запятой на узел сетки или сложное обобщение на упругую среду.

Другим методом постобработки является подход глубокого обучения, который может оказаться наиболее продуктивным из-за его универсальности в обработке данных. Нейронная сеть для уменьшения численной дисперсии (NDM-net) была впервые предложена в [3], а затем разработана в [4,5].

Однако эффективность подходов глубокого обучения в значительной степени зависит от репрезентативной выборки, на которой будет обучаться нейронная сеть. Таким образом, в статье поднимается проблема формирования обучающей выборки как наиболее важной. Недавно мы предложили использовать три показателя для анализа сейсмограмм: расстояние между источниками [3], расстояние между сейсмограммами [5] и расстояние между скоростными моделями [4]. Мы использовали кластерный анализ для построения репрезентативной выборки. Таким образом, мы учли свойства самих данных и с помощью такого подхода смогли сократить обучающую выборку без потери качества обучения.

Нейронная сеть

Для подавления численной дисперсии с помощью методов машинного обучения предлагается следующий алгоритм:

1. Рассчитать сейсмограммы $\vec{u}_{h_2}^k$ с использованием грубой вычислительной сетки с характерным шагом h_2 для всех положений источников $x_s^k, k = 1, \dots, N_s$;

2. Сформировать набор индексов источников $J_t \subset \{1, \dots, N_s\}$;

3. Рассчитать сейсмограммы $\vec{u}_{h_1}^k, k \in J$ с использованием мелкой расчётной сетки с шагом $h_1 < h_2$;

4. Обучить нейронную сеть $G: \vec{u}_{h_2}^k \to \vec{u}_{h_1}^k$ так, чтобы для всех $k \in J$ выполнялось следующее условие:

$$\left\| G\left(\vec{u}_{h_2}^k\right) - \vec{u}_{h_1}^k \right\| \ll \left\| \vec{u}_{h_2}^k - \vec{u}_{h_1}^k \right\|,$$

где $G -$ оператор перехода из данных, рассчитанных на грубой сетке, в данные, рассчитанные на мелкой сетке;

5. Применить обученную нейронную сеть на весь набор сейсмограмм:

$$\vec{u}_G^k = G\left(\vec{u}_{h_2}^k\right),$$

для всех $k \in \{1, \dots, N_s\}$.

Способы формирования обучающей выборки

В работе обсуждаются три способа формирования выборки. Все они основаны на иерархическом кластерном анализе, который использует матрицу расстояний. Так, вводится три расстояния, которые отражают свойства данных. Первое расстояние – это евклидово расстояние между положениями источников:

$$d_d^{ij} = \sqrt{\left(r_s^i - r_s^j\right)^2},$$

где $r_s^i -$ это положение i −ого источника.

Второе расстояние – это прямое расхождение сейсмограмм на основе $L_2 -$ нормы:

$$d_s^{ij} = 2\frac{\left\|\vec{u}^i - \vec{u}^j\right\|_2}{\|\vec{u}^i\|_2 + \|\vec{u}^j\|_2}.$$

Третье расстояние – это расстояние между скоростными моделями, которое выглядит следующим образом:

$$d_m^{ij} = 2\frac{\left\|M^i - M^j\right\|_2}{\|M^i\|_2 + \|M^j\|_2},$$

где $M -$ это скоростная модель, описываемая как

$$\left\|M^i\right\|_2^2 = \int_0^Z \int_{x_s^i - L_x}^{x_s^i + L_x} \left(v_p^2(x,z) + v_s^2(x,z)\right) dxdz\,,$$

где $Z, L_x -$ глубина модели и максимальное смещение в расчётной области $D = [x_s - L_x, x_s + L_x] \times [0, Z]$, $v_p -$ скорость продольной волны, определяемая как $v_p = \sqrt{\frac{\lambda+2\mu}{\rho}}$, $v_s -$ скорость поперечной волны, определяемой как $v_s = \sqrt{\frac{\mu}{\rho}}$.

Далее, введём Хаусдорфовы расстояния для подсчёта расстояния между кластерами, а также для статистического анализа, который будет описан позже:

$$\delta_d = \max_{x_s^i \in S} \min_{x_s^j \in S_0} d_d^{ij}\left(x_s^i, x_s^j\right), \qquad \delta_s = \max_{u^i \in S} \min_{u^j \in S_0} d_s^{ij}, \qquad \delta_m = \max_{M^i \in S} \min_{M^j \in S_0} d_d^{ij}$$

где S −набор сейсмограмм, включенных в обучающую выборку, $S_0 -$ набор всех сейсмограмм.

Оптимизированная выборка

Для оценки влияния каждой из введённых метрик на качество обучающей выборки воспользуемся статистическим анализом, а, именно, глобальным анализом чувствительности. Для этого формируются случайным образом обучающие выборки и обучаются статистически значимое число NDM-net. Зафиксируем размер обучающей выборки $N_t = const$ и рассмотрим среднюю ошибку выхода нейронной сети относительно точного решения для каждого N_c в виде функции, зависящей от трёх метрик:

$$\varepsilon_{N_c} = \varepsilon_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_3\delta_3 + o(\delta_1, \delta_2, \delta_3),$$

где $\varepsilon_{N_c} = \frac{1}{N_c}\sum_{i=1}^{N_c} d_s(G(\vec{u}_{h_2}), \vec{u}_{h_1})$, $(\delta_1, \delta_2, \delta_3) = (\delta_d, \delta_s, \delta_m)$.

Минимизаця такой ошибки эквивалентная минимизации линейной комбинации:

$$l(\delta_1, \delta_2, \delta_3) = q(\alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_3\delta_3) \to min, \qquad \forall q > 0.$$

Коэффициенты $\alpha_1, \alpha_2, \alpha_3$ определяются из регрессивного анализа и затем используются для расчёта коэффициентов Соболя, которрые для линейной функции будут выглядеть следующим образом:

$$S_{\delta_i} = \frac{Var\left(\mathbb{E}_{\vec{\delta}\sim i}(\varepsilon|\delta_i)\right)}{Var(\varepsilon)} = \frac{\left(\alpha_i L_{\delta_i}\right)^2}{\sum_j \left(\alpha_j L_{\delta_j}\right)^2}, \qquad S_{\delta_i} \in [0,1], \sum_i S_{\delta_i} = 1, i = 1,2,3,$$

где $\mathbb{E}$ −математическое ожидание; Var −вариация; $\vec{\delta}\sim i$ − это вектор всех параметров за исключением i − ого. Индексы Соболя первого порядка иллюстрируют, насколько сильна изменчивость рассматриваемой функции по отношению к одному входному параметру. Для формирования новой метрики, оцениваются коэффициенты:

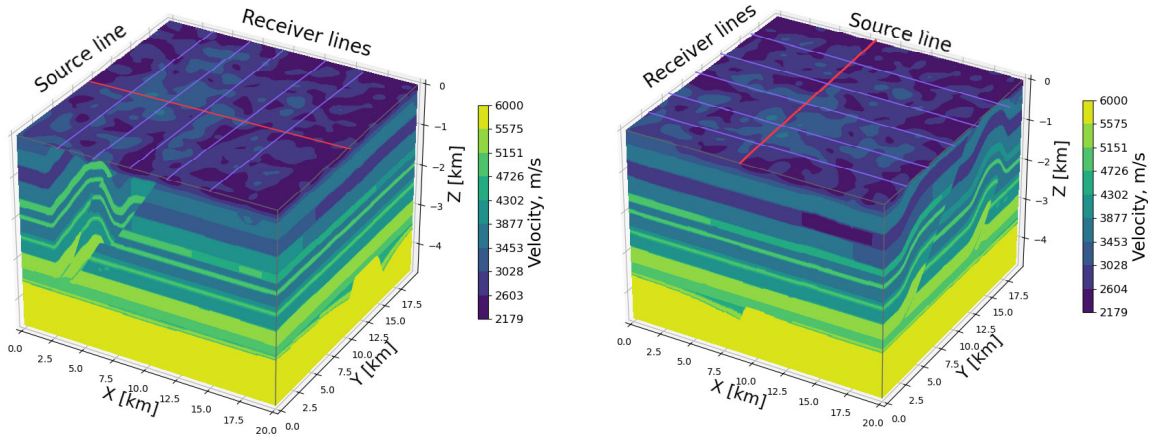
$$\alpha_i = \frac{1}{N_r}\sum_{k=0}^{N_r} \frac{\sqrt{S_i^k \sum_j \left(\alpha_j^k L_j^k\right)^2}}{L_i^k},$$

где N_r −количество рандомных тренировочных датасетов, L_i^k − разброс δ_i.

Входные данные

Для проведения численных экспериментов в качестве синтетической трехмерной модели была использована модель Overthrust, имеющая разлом, уровень

соли в основании и слоистую структуру. Размер модели составляет $20 \times 20 \times 4.675$ км^3. Используется система сбора данных, которая имитирует 3D-моделирование, то есть линии приема перпендикулярны линиям источника, как показано на рис. 1. Линии приема расположены на расстоянии 20 м друг от друга, в то время как линии источников расположены на расстоянии 100 м друг от друга. Количество источников в одной линии – 1001. Количество приёмников для каждого источника – 513. В общей сложности, в наборе данных 401401 источников.



а) *б*)

Рис. 1. Overthrust скоростная модель и система наблюдений:
a) crossline; b) inline

Мы рассчитываем сейсмограммы с шагом в 5 и 2,5 м, расчёт ведётся до 5 с и шагом по времени в 2 мс. В качестве источника используется импульс Рикеса с центральной частотой 30 Гц.

Результаты

Для статистического анализа были сгенерированы около 1000 случайных выборок размером от 0.5 % до 2% от общего количества. После обучения NDM-net на каждом таком обучающем наборе, была рассчитана средняя ошибка между всем набором точных сейсмограмм и набором сгенерированных нейронной сетью сейсмограмм. Оценки коэффициентов линейной части функции ошибки вплоть до скалярного множителя:

$$\alpha_d^{Overthrust} = 0.37, \alpha_s^{Overthrust} = 0.33, \alpha_m^{Overthrust} = 0.3.$$

Так, можно сделать вывод о том, что все параметры для трёхмерной среды являются в равной степени значимыми и можно ввести новую метрику в виде комбинаций трёх ранее введённых метрик следующим образом:

$$0.37\delta_d + 0.33\delta_s + 0.3\delta_m \to min$$

Таким образом, градиент выходной ошибки аппроксимируется линейной комбинацией трёх введённых метрик, сводя задачу к минимизации по одному параметру. На рис. 2 построен график ошибок, соответствующих минимизации только одной метрики (D^{new}), и для случайно сгенерированных наборов данных (D^{random}), чтобы проиллюстрировать, что оптимизированный набор данных обеспечивает, в целом, меньшую ошибку, чем любой другой ранее рассмотренный способ генерации обучающего набора данных. Значение изначальной средней ошибки составляло 0.15.

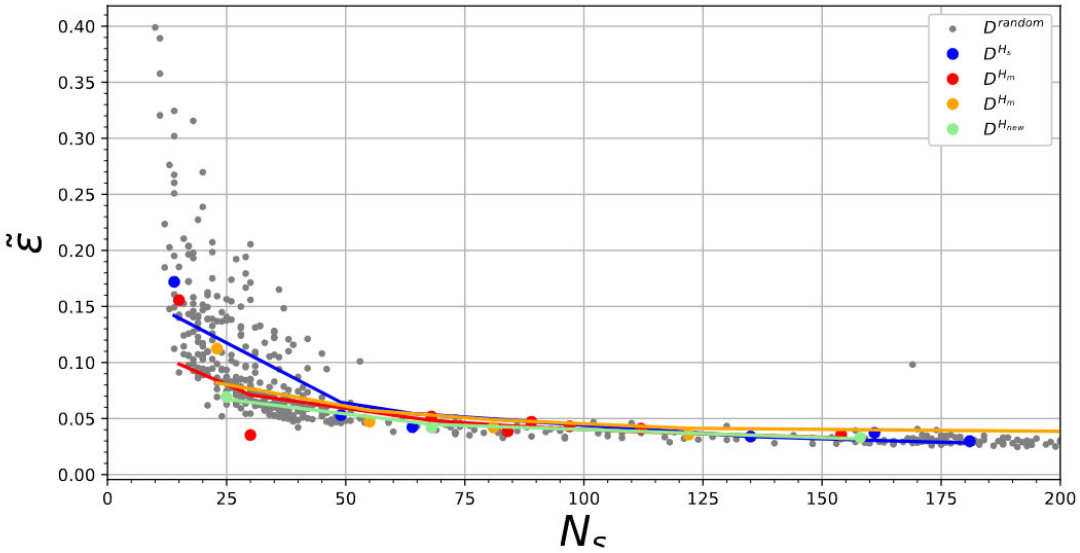


Рис. 2. Зависимость средней выходной ошибки NDM-net ко всему набору данных от размера обучающей выборки для разных типов выборки для Overthrust модели

Заключение

В этом исследовании мы разработали метод уменьшения числовой дисперсии в сейсмических данных с использованием системы 3D-сбора данных, сосредоточив внимание на построении репрезентативной выборки. Предлагаемый метод включает в себя расчет матрицы расстояний с использованием трех различных показателей, отражающих свойства данных, и кластерный анализ. Мы провели численные эксперименты на синтетических данных и показали, что все методы построения хорошо обобщаются на трехмерный случай. Предлагаемый метод позволяет снизить численное отклонение от первоначального среднего значения в 3-4 раза в используемой норме.

Благодарности

Авторы выражают благодарность за финансовую поддержку РНФ № 22-11-00004.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Liu Y., Sen M. K. A new time–space domain high-order finite-difference method for the acoustic wave equation // Journal of computational Physics. – 2009. – Vol. 228. – № 23. – P. 8779-8806.

2. Levander A. R. Fourth-order finite-difference P-SV seismograms // Geophysics. – 1988. – Vol. 53. – № 11. – P. 1425-1436.

3. Gadylshin K., Vishnevsky D., Gadylshina K., Lisitsa V. Numerical dispersion mitigation neural network for seismic modeling // Geophysics. – 2022. – Vol. 87. – № 3. – P. T237-T249.

4. Gondyul E., Lisitsa V., Gadylshin K., Vishnevsky D. Numerical dispersion mitigation neural network with the model-based training dataset optimization // International Conference on Computational Science and Its Applications. – Cham : Springer Nature Switzerland, 2023. – P. 19-30.

5. Gadylshin K., Lisitsa V., Vishnevsky D., Gadylshina K. Hausdorff-distance-based training dataset construction for numerical dispersion mitigation neural network // Computers and Geosciences. – 2023. – Vol. 180. – P. 105438.