ТЕОРЕТИЧЕСКАЯ И МАТЕМАТИЧЕСКАЯ ФИЗИКА

Активационные функции для глубокого обучения, построенные на основе обобщённых энтропий

Р.А. Рудаменко,[1,*] А.М. Савченко,[1,†] К.М. Семенов[1,‡]

[1] *Московский государственный университет имени М. В. Ломоносова, физический факультет. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2*

(Поступила в редакцию 28.01.2026; подписана в печать 11.02.2026)

Энтропия Шеннона (Больцмана–Гиббса) является фундаментом классической статистической механики и глубокого обучения, однако она имеет сложности с описанием динамики неэкстенсивных систем. В данной работе предлагается применение обобщённых энтропий для построения новых фундаментальных блоков в архитектурах глубоких нейронных сетей. Предложенный подход обобщает классический слой softmax с использованием параметрических энтропий Реньи, Тсаллиса и Шарма–Миттала. Параметры q и r контролируют форму распределения: при $q \to 1$ оптимальное распределение сходится к softmax, а при $q = 2$ — к sparsemax. В частности, рассматривается вариант, соответствующий q-entmax, где адаптивность достигается изменением параметра q при фиксированном r. Исследование включает получение аналитических выражений якобиана по параметрам q и r с целью оптимизации через методы явного дифференцирования. Проведен сравнительный анализ с существующими подходами — softmax, sparsemax и entmax (при $q \in \{1.25, 1.5, 1.75\}$). Полученные результаты демонстрируют рост метрик качества относительно подходов с softmax, sparsemax и q-entmax для задачи классификации со скоррелированными метками классов, что позволяет сделать вывод о преимуществе метода на основе Шарма–Миттала для поставленной задачи.

PACS: 05.20.-y, 05.70.-a, 05.90.+m. УДК: 536.7

Ключевые слова: обобщённые распределения, энтропия Шарма–Миттала, глубокое обучение.

DOI: 10.55959/MSU0579-9392.81.2620101

ВВЕДЕНИЕ

Классическая энтропия Шеннона, совпадающая с термодинамическим подходом Больцмана–Гиббса, стала прочной основой критерия *кросс-энтропия* и нормировки `softmax`, переводящей выходы глубокой сети в плотное распределение вероятностей (не содержащее большое количество нулей). При всём элегантном минимализме эта мера не описывает статистику систем с дальнодействием, степенными хвостами и фрактальной матрицей состояний [1–3]. В нейронных сетях такое распределение ведёт к размыванию градиентов и потере интерпретируемости. Принцип максимума энтропии Джейнса [4] и геометрическая теория информации Амари [5] сформировали язык, на котором удобно говорить о деформациях энтропии. В этом языке работы Реньи [2, 6], Тсаллиса[1] и, шире всего, Шарма–Миттала [6, 7] используют дополнительную параметризацию q, r, позволяющую плавно переходить от экспоненциальных к степенным распределениям. Обзор [8] отмечает характерное увеличение таких подходов в современном анализе данных.

Приложения параметрических энтропий находят применение в широком множестве дисциплин: в исследованиях лечения мутаций энтропия Тсаллиса повышает точность стратификации мутаций по экспрессии генов [8, 9], а в анализе медицинских изображений применение q-деформированных энтропий улучшает сегментацию фрактальных контуров опухолей на МРТ и КТ [8]. В эпидемиологии сложных сетей оценка устойчивости распространения инфекций на гетерогенных графах через энтропию Тсаллиса позволяет обнаруживать нетривиальные пороги заражения [8, 10]. В квантовых технологиях энтропия Реньи второго порядка (с $q = 2$) применяется для измерения запутанности и валидации квантового превосходства [11, 12], тогда как в задачах цифровой безопасности дивергенция на основе энтропии Тсаллиса выявляет ботнеты и DGA-домены с большей чувствительностью, чем дивергенция Кульбака–Лейблера [13]. Наконец, в финансах q-энтропийная риск-мера даёт устойчивые оценки волатильности портфеля на негауссовых рынках [8, 14], а в космологии энтропия Шарма–Миттала используется для описания термодинамики чёрных дыр и аномального расширения Вселенной [8, 15]. Эти примеры демонстрируют универсальность адаптации параметров q, r для широкого спектра задач, фактически задавая геометрию и форму оптимального распределения, согласуемую с наблюдаемой статистикой и требованиями задачи.

В данной работе мы получаем слой `SharMiX` как дифференцируемое отображение

* E-mail:rudamenk@gmail.com
† E-mail:a.m.savchenko@gmail.com
‡ E-mail:semenovkm@protonmail.com

$\mathbf{p} = \texttt{SharMiX}(\mathbf{z}; q, r)$ из вектора выдаваемых оценок моделью $\mathbf{z}$ для задачи классификации в вероятностное распределение на симплексе, вычисляем его якобиан по $\mathbf{z}$ для обратного распространения ошибки, а также производные по параметрам (q, r) [16, 17]. Лучшая конфигурация `SharMiX` достигает Accuracy 0.9146 и Macro-F1 0.8863, превосходя softmax и sparsemax при идентичности установки, а также существенно опережая entmax при разных значениях параметра q. Это подчёркивает значимость использования `SharMiX` для рассматриваемой задачи. Для *энергетической и ресурсной эффективности* разреженные трансформации `sparsemax` [18] и q-`entmax` [19] снижают вычислительную сложность операций и энергопотребление TPU-кластеров [20]. Рассматриваемый в работе подход `SharMiX`, используя два независимых параметра, обещает ещё более впечатляющие результаты.

Уточнение аналогий между подходами. Softmax (Гиббс-нормировка) даёт распределение с полным носителем, при котором статистический вес каждой компоненты строго положителен при конечной температуре; *sparsemax* заменяет экспоненциальную нормировку на проекцию на симплекс, вследствие чего часть компонент попадает на границу и получает строго нулевой вес (носитель распределения сокращается, «хвост» отсекается); q-entmax задаёт однопараметрическую траекторию между этими режимами и для $q > 1$ порождает разреженные распределения. Двухпараметрическая трансформация на основе Шарма–Миттала способствует новому взгляду на предыдущие подходы, но до сих пор не имела удобных градиентов.

В разделе 1 описывается формализм энтропий Реньи и Шарма–Миттала, доказывается эквивалентность якобианов и вычисляется распределение задачи оптимизации. Затем в разделе 2 мы отмечаем аппаратные и вычислительные тонкости эксперимента. Результаты продемонстрированы в разделе 3.

Таким образом, в данной работе мы показываем, что обобщение методов глубокого обучения на основе теории двухпараметрических энтропий демонстрирует заметное улучшение результатов в эксперименте.

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОБОБЩЁННЫХ ЭНТРОПИЙ

1.1. Интерпретация через свободную энергию

В духе информационного подхода Джейнса энтропийный критерий можно трактовать как меру неполноты знания о системе. Вычислительно эта неопределённость уменьшается оптимизацией параметров модели методом градиентного спуска по выбранному функционалу потерь. В статистической физике критерий максимума энтропии тесно связан с понятием свободной энергии. В частности, известно, что относительная энтропия Кульбака–Лейблера D_{KL} между произвольным распределением p_i и равновесным распределением $\tilde{p}_i$ эквивалентна разности свободных энергий этих состояний (с точностью до температуры) [21, 22]. Свободная энергия задаётся выражением

$$F = U - TS, \tag{1}$$

где $U = \sum_i p_i \varepsilon_i$ — внутренняя энергия (среднее значение энергии состояний ε_i), а S — энтропия системы. Минимизация свободной энергии F из формулы (1) при фиксированной температуре T приводит к равновесному распределению Больцмана–Гиббса $p_i^{(\mathrm{eq})} \propto \exp(-\varepsilon_i/T)$ и при этом

$$T\, D_{\mathrm{KL}}(p\|\tilde{p}) = F[p] - F[\tilde{p}], \tag{2}$$

то есть D_{KL} выражает выигрыш (уменьшение) свободной энергии согласно (2). Таким образом, стандартная энтропия Шеннона выступает мерой «отдалённости» состояния от равновесия через призму избыточной свободной энергии.

Обобщённый функционал Шарма–Миттала также можно рассматривать с этой физической точки зрения. Его энтропия $H_{r,q}(p)$ вводится как двухпараметрическое семейство (см. п. 2.3) и включает в качестве предельных случаев энтропии Реньи и Тсаллиса. Функция свободной энергии в таком случае может быть задана как

$$\mathcal{F}_{SM}[p] \;=\; U[p] \;-\; T\, H_{r,q}(p). \tag{3}$$

В выражении (3) параметр T задаёт относительный вес энтропийного члена, как и в каноническом ансамбле: при увеличении T минимум $\mathcal{F}_{SM}$ смещается в сторону более «размазанного» распределения. Параметры q и r, в свою очередь, деформируют *саму* энтропию $H_{r,q}$, т. е. меняют кривизну и форму функционала $\mathcal{F}_{SM}[p]$ как «ландшафта» над множеством распределений p; это приводит к изменению положения и геометрии минимума даже при фиксированном T. Такое деформирование действует в вариационной постановке аналогично изменению температуры в выражении соотношения «энергия–энтропия», однако параметр q не является температурой и не предполагается функцией T: в неэкстенсивной статистике q рассматривается как независимый параметр, связанный с неаддитивностью/корреляциями и (в ряде моделей) с флуктуациями интенсивных величин, в частности температуры [23–25]. Экстремум $\mathcal{F}_{SM}$ по p при заданных ограничениях определяет *обобщённое равновесие*, которое характеризуется распределением, максимизирующем энтропию Шарма–Миттала при фиксированном U.

В общем случае функционал $\mathcal{F}_{SM}$ можно трактовать как *обобщённую свободную энергию* системы, и разность $\mathcal{F}_{SM}[p] - \mathcal{F}_{SM}[p^{(\mathrm{eq})}]$ характеризует отклонение p от обобщённого термодинамического равновесия. Однако важным нюансом является выбор параметров q, r: в пределе $q, r \to 1$ энтропия

Шарма–Миттала сводится к классической и её относительная энтропия действительно равна разности свободных энергий между начальной инициализации системы и обновлённой её копией с помощью градиентного спуска. Вне этого предела прямая интерпретация относительной энтропии Шарма–Миттала как разности свободных энергий требует осторожного обращения — дополнительные члены, связанные с неэкстенсивностью, не позволяют столь же просто представить функционал как разность F двух состояний.

Тем не менее, сами термодинамические величины в обобщённом формализме сохраняют аналоги: вводятся параметризация энергии параметрами r и q, энтропия $H_{r,q}$ и сопряжённая температура (часто обозначаемая Θ), так что справедливо соотношение

$$d\mathcal{F}_{SM} = -\Theta\, dH_{r,q} + dU, \tag{4}$$

где U — работа поля внешних сил, оптимизирующих систему.

Подобно случаю Реньи, для которого показано соответствие между энтропией и свободной энергией системы при изменении температуры [2], функционал Шарма–Миттала можно рассматривать как свободную энергию эффективной системы с «деформированной» статистикой. Формулируя в терминах нейронной сети, *логиты* z_i, используемые в модели (например, выходы нейронной сети до softmax-нормировки), аналогично энергии ε_i определяют вероятность $p_i \sim \exp_{q,r}(-\varepsilon_i/T)$ через обобщённую экспоненту. В этом смысле оптимизация с учетом обобщённой энтропии эквивалентна достижению баланса между энергией и энтропией в неэкстенсивной термодинамике системы.

Такой физический взгляд помогает интерпретировать поведение модели. В каноническом формализме уменьшение температуры T делает распределение Больцмана–Гиббса более концентрированным на низкоэнергетических состояниях. В нашей постановке параметр T аналогично задаёт относительный вес энтропийного вклада. Параметры q и r действуют иначе: они деформируют сам энтропийный функционал $H_{r,q}$ и тем самым меняют кривизну и форму функционального «ландшафта» $\mathcal{F}_{SM}[p]$ на симплексе вероятностей, что приводит к смещению минимума и изменению концентрации/разреженности оптимального распределения даже при фиксированном T. При значениях $q \to 1$ восстанавливается больцмановско–гиббсовский случай (softmax), тогда как при движении q в сторону 2 (в рамках нашей модели) реализуется режим, близкий к `sparsemax` (с сокращением носителя). Поэтому изменение q можно трактовать как «эффективное охлаждение» ландшафта в смысле усиления реализации наиболее вероятных компонент, так как q не является температурой и не предполагается функцией T: в неэкстенсивной статистике q рассматривается как независимый индекс, который в ряде моделей связывают с флуктуациями интенсивного параметра (например, β) в рамках суперстатистики [24].

1.2. Эквивалентность якобианов для энтропий Тсаллиса и Реньи

В нашей статье мы решаем задачу максимизации энтропии на симплексе:

$$\max_{p\in\Delta^N} \Big\{\mathbf{p}^\top \mathbf{z} + H(p)\Big\}, \tag{5}$$

где $\Delta^N = \{\mathbf{p} \in \mathbb{R}^N : p_i \geq 0,\ \sum_i p_i = 1\}$ и $H(p)$ могут быть выбраны как энтропия Реньи или Тсаллиса.

Отметим, что энтропия Тсаллиса задается следующим выражением:

$$H_q^{\text{Tsallis}}(p) = \frac{1}{q-1}\Big(1 - \sum_{i=1}^N p_i^q\Big) \tag{6}$$

и в реализации функции *entmax* записывается как

$$H_q^{\text{Tsallis}}(p) = \frac{1}{q(q-1)} \sum_{i=1}^N \Big(p_i - p_i^q\Big), \tag{7}$$

что эквивалентно стаднартному определению с точностью до множителя.

Энтропия Реньи задаётся выражением

$$H_q^{\text{Rényi}}(p) = \frac{1}{1-q} \ln\Big(\sum_{i=1}^N p_i^q\Big). \tag{8}$$

Поскольку логарифм является монотонной функцией, оптимизация задачи

$$\max_{p\in\Delta^N} \Big\{\mathbf{p}^\top \mathbf{z} + H(p)\Big\} \tag{9}$$

при использовании $H_q^{\text{Tsallis}}(p)$ или $H_q^{\text{Rényi}}(p)$ приводит к одному и тому же оптимальному распределению. В частности, условия оптимальности Каруша–Куна–Такера [26–28] для обеих энтропий приводят к соотношению

$$z_i + \frac{\partial}{\partial p_i} H(p) = \lambda, \quad \forall\, i \text{ с } p_i > 0, \tag{10}$$

что после явного вычисления производной дает (с точностью до множителя):

$$p_i^\star \propto [\, z_i - \tau\,]_+^{\frac{1}{q-1}}, \tag{11}$$

где $[x]_+ = \max\{x, 0\}$ и τ — порог, обеспечивающий нормировку $\sum_i p_i = 1$.

Таким образом, оптимальное распределение в нашем подходе имеет вид:

$$p^\star = \frac{[\, z - \tau\,]_+^{\frac{1}{q-1}}}{\sum_j [\, z_j - \tau\,]_+^{\frac{1}{q-1}}}, \tag{12}$$

независимо от того, используется ли энтропия Тсаллиса или Реньи. При вычислении градиентов по выходу последнего слоя глубокой сети z для этого оптимального распределения получаются одинаковые

якобианы с точностью до масштабирующих постоянных.

Эти результаты подтверждаются в работах [16, 29], где подробно анализируется связь между обобщёнными энтропиями Тсаллиса и Реньи, а также в классической статье [1] и в обзоре [6]. Кроме того, анализ энтропии Шарма–Миттала [7] показывает, что при выборе параметров r и q (в обозначениях Masi [30]) энтропия Шарма–Миттала может перейти в энтропию Реньи, что дополнительно подтверждает эквивалентность оптимальных распределений и их градиентных свойств.

Таким образом, для задач, где применяется оптимизация вида

$$\mathbf{p}^\top \mathbf{z} + H(p), \tag{13}$$

выбор между энтропией Тсаллиса и энтропией Реньи не влияет на оптимальное распределение, а следовательно, и на форму якобиана по выходу последнего слоя глубокой сети. Это позволяет использовать любую из мер в зависимости от удобства реализации, что особенно важно для построения разреженных активационных слоёв в глубоких нейронных сетях [16, 19].

1.3. Энтропия Шарма–Миттала и оптимизация параметров q и r

Энтропия Шарма–Миттала представляет собой двухпараметрическое обобщение функционалов энтропий Реньи и Тсаллиса и задаётся выражением

$$H_{r,q}(p) = \frac{1}{1-r}\left[\left(\sum_{i=1}^{N} p_i^q\right)^{\frac{1-r}{1-q}} - 1\right], \tag{14}$$

где q и r — параметры, контролирующие форму энтропии согласно статье [7]. При $r \to 1$ энтропия Шарма–Миттала сходится к энтропии Реньи, а при $r \to q$ — к энтропии Тсаллиса. Это обобщение позволяет гибко регулировать форму распределения, получаемого в результате максимума энтропии, поскольку изменение параметра q ведёт к переходу от softmax (при $q \to 1$) к sparsemax (при $q = 2$) [29].

Задача оптимизации в данном случае имеет вид

$$\max_{p \in \Delta^N} \left\{\mathbf{p}^\top \mathbf{z} + \gamma\, H_{r,q}(p)\right\}, \tag{15}$$

где γ — коэффициент регуляризации. Анализ условий оптимальности с использованием метода множителей Лагранжа приводит к тому, что оптимальное распределение имеет тот же вид:

$$p_i = \frac{[z_i - \tau]_+^{\frac{1}{q-1}}}{\sum_{j=1}^{N} [z_j - \tau]_+^{\frac{1}{q-1}}}, \tag{16}$$

но порог τ теперь неявно зависит от обоих параметров q и r. Для вычисления градиентов по параметрам, таких как $\frac{\partial p}{\partial q}$ и $\frac{\partial p}{\partial r}$, можно использовать метод неявного дифференцирования уравнения нормировки

$$f(\tau, q, r, z) = \sum_{i=1}^{N} p_i(\tau, q, r, z) - 1 = 0, \tag{17}$$

что позволяет получить аналитические выражения якобиана. Эти выражения обеспечивают возможность оптимизации параметров q и r с использованием стандартных оптимизаторов первого порядка, таких как Adam [31], и второго, таких, как K-FAC [32] и LBFGS [33] при условии, что параметры заданы как обучаемые и корректно ограничены (например, с помощью параметризации) [16, 17].

Таким образом, несмотря на различия в исходных определениях, энтропия Реньи имеет логарифмический вид, а энтропия Тсаллиса — степенной, их оптимальные решения в задаче максимума энтропии совпадают (с точностью до постоянных коэффициентов). Это приводит к тому, что градиенты по входам z (якобианы) для обеих мер эквивалентны, что подтверждается аналитически и экспериментально [19, 29]. Дополнительное обобщение в виде энтропии Шарма–Миттала позволяет еще гибче настраивать форму распределения, что особенно полезно в сложных системах, где требуется оптимальность выхода.

1.4. Вывод якобиана по параметру q для параметрической энтропии

Рассмотрим вывод на примере энтропии Реньи. Этот результат будет совпадать и для случаев Тсаллиса и Шарма–Миттала (при варьировании параметра r последнего).

Пусть оптимальное распределение, полученное в результате решения задачи

$$\max_{\mathbf{p} \in \Delta^N} \left\{\mathbf{p}^\top \mathbf{z} + \lambda\, H_q^{\text{Rényi}}(\mathbf{p})\right\}, \tag{18}$$

имеет вид

$$p_i^*(q) = \frac{[(q-1)(z_i - \tau(q))]_+^{\frac{1}{q-1}}}{\sum_{j=1}^{N} [(q-1)(z_j - \tau(q))]_+^{\frac{1}{q-1}}}, \tag{19}$$

где $[x]_+ = \max\{x, 0\}$ и пороговое значение, обеспечивающее $\tau(q)$ нормировку:

$$\sum_{i=1}^{N} p_i^*(q) = 1. \tag{20}$$

Для удобства введём обозначение:

$$F(q, z_i, \tau(q)) \;=\; \frac{1}{q-1} \log\left[(q-1)(z_i - \tau(q))\right], \tag{21}$$

так что можно записать:

$$p_i^*(q) \;=\; \exp\Big(F(q, z_i, \tau(q))\Big). \tag{22}$$

Чтобы получить производную $\frac{\partial p_i^*}{\partial q}$, применим правило дифференцирования экспоненты:

$$\frac{\partial p_i^*}{\partial q} = p_i^*(q) \cdot \frac{\partial F}{\partial q}. \quad (23)$$

Найдем теперь $\frac{\partial F}{\partial q}$. Поскольку

$$F(q, z_i, \tau(q)) = \frac{1}{q-1} \log\Big[(q-1)(z_i - \tau(q))\Big], \quad (24)$$

получаем, используя правила дифференцирования частного и правило дифференцирования сложной функции,

$$\frac{\partial F}{\partial q} = -\frac{1}{(q-1)^2} \log\Big[(q-1)(z_i - \tau(q))\Big] + \\ + \frac{1}{q-1} \cdot \frac{1}{(q-1)(z_i - \tau(q))} \times \\ \times \left[(z_i - \tau(q)) - (q-1)\frac{\partial \tau(q)}{\partial q} \right]. \quad (25)$$

Упростим полученное выражение:

$$\frac{\partial F}{\partial q} = \frac{1}{(q-1)^2} \times \\ \times \Big[\frac{z_i - \tau(q) - (q-1)\tau'(q)}{z_i - \tau(q)} - \log\Big((q-1)(z_i - \tau(q))\Big) \Big], \quad (26)$$

где мы обозначили $\tau'(q) = \frac{\partial \tau(q)}{\partial q}$.

Окончательное выражение для производной оптимального распределения по параметру q для тех случаев, где $p_i^*(q) > 0$, имеет вид:

$$\frac{\partial p_i^*}{\partial q} = p_i^*(q) \cdot \frac{1}{(q-1)^2} \Big[\frac{z_i - \tau(q) - (q-1)\tau'(q)}{z_i - \tau(q)} - \\ - \log\Big((q-1)(z_i - \tau(q))\Big) \Big]. \quad (27)$$

При $p_i^*(q) = 0$ производная принимается равной нулю.

Значение $\tau'(q)$ можно определить через неявное дифференцирование уравнения нормировки:

$$f(q, \tau(q)) = \sum_{i=1}^{N} p_i^*(q) - 1 = 0. \quad (28)$$

Дифференцируя это уравнение по q, получаем

$$\frac{\partial f}{\partial q} + \frac{\partial f}{\partial \tau} \tau'(q) = 0, \quad (29)$$

и, соответственно,

$$\tau'(q) = -\frac{\frac{\partial f}{\partial q}}{\frac{\partial f}{\partial \tau}}. \quad (30)$$

Аналитическое выражение для $\tau'(q)$, как правило, имеет сложную форму, его можно вычислять либо аналитически, либо приближённо с использованием методов неявного дифференцирования [16, 29].

Таким образом, полученное выражение для $\frac{\partial p_i^*}{\partial q}$ демонстрирует, что якобиан по параметру q зависит как от значений выхода последнего слоя глубокой сети z_i и оптимального порога $\tau(q)$, так и от неявной зависимости $\tau'(q)$. Это позволяет корректно распространять градиенты по параметру q в задачах обучения, где q является обучаемым параметром адаптивных активаций (например, в контексте q-entmax) [17, 19].

Замечание. Следует отметить, что данное выражение имеет корректное поведение в пределе $q \to 1$ (когда энтропия Реньи переходит в энтропию Шеннона) и обеспечивает непрерывность обратного распространения, несмотря на неявную зависимость $\tau(q)$.

1.5. Ландшафты оптимизации и аналогия со спиновыми стёклами

Минимизируемые функции в машинном обучении часто обладают множеством локальных экстремумов. Это роднит их с энергическими ландшафтами спиновых стёкол — физических систем с беспорядочно взаимодействующими спинами. *Спиновое стекло* характеризуется «шероховатым» ландшафтом свободной энергии с бесчисленным множеством локальных минимумов, разделённых высокими барьерами. Система в спиновом стекле не может легко перейти из одного минимума в глобальный минимум из-за метастабильности: существуют долгоживущие состояния, в которых система «застревает».

В формализме статистической механики такое поведение описывается методом реплик и *нарушением симметрии реплик* [34, 35]. Паризи показал [36–38], что для полного описания равновесного состояния спинового стекла требуется иерархия реплик, характеризующая распределение перекрытий между различными метастабильными состояниями. Множество локальных минимумов («чистых состояний») образует иерархическую структуру, где разные решения модели (конфигурации спинов или параметры нейросети) слабо связаны друг с другом. Каждое метастабильное состояние можно рассматривать как экстремальную компоненту равновесной смеси, и нарушение симметрии реплик формализует статистическое множество таких состояний. Для того, чтобы отличать метастабильное состояние от стабильного, вводится понятие *сложности ландшафта* — функции $\Sigma(f)$, показывающей логарифмическую плотность числа локальных экстремумов с данным уровнем свободной энергии f. В спиновых стёклах $\Sigma(f) > 0$ в стеклянной фазе, что указывает на экспоненциально большое число локальных минимумов и называется мультиэкстремальностью ландшафта.

Важно подчеркнуть, что аналогичные явления могут возникать при оптимизации функционала Шарма–Миттала в сложных моделях. Неэкстен-

сивность и дополнительная параметризация (q, r) приводят к тому, что функция потерь становится негладкой. При некоторых значениях q (например, при $q > 1$) энтропийный член может делать функцию *не*вогнутой, открывая возможность множества локальных максимумов оптимизационной цели. В результате поиск оптимума распределения или параметров модели сталкивается с «шероховатым» ландшафтом: присутствуют многочисленные стационарные точки — локальные минимумы и сёдла, разделённые барьерами по функции потерь, подобно фазе спинового стекла.

Метастабильность в таком ландшафте проявляется как чувствительность алгоритма оптимизации к начальной инициализации: различные запуски градиентного спуска могут приводить к разным локальным решениям, аналогично тому, как спиновое стекло при быстром охлаждении может попадать в разные застеклованные состояния.

Заметим, что в контексте глубокого обучения подобные аналогии уже проводились. Например, в работе [39] было показано, что при увеличении размерности сети динамика обобщения испытывает фазовый переход от «эргодичного» режима к режиму с замороженной ошибкой, напоминающий переход к спиново-стекольному состоянию. В недавней работе [40] обсуждается, как широкий минимум функции ошибки (flat minimum) можно интерпретировать как область с большим числом эквивалентных решений, аналогично вырожденным состояниям в спиновом стекле. В целом ландшафт обучения нейросети может обладать «стекольными» чертами и дополнительная энтропийная регуляризация (например, внедрение функции Шарма–Миттала) способна как сглаживать ландшафт, так и при неудачном выборе параметров q, r вводить новые локальные минимумы.

Следует подчеркнуть, что оптимизация с функционалом Шарма–Миттала индуцирует эффект нескольких минимумов, сходный со спиновым стеклом, когда параметры q, r уходят от экстенсивного случая. Так, при больших q выходное распределение модели становится более *разреженным* (sparse), отдавая нулевую вероятность некоторым классам подобно «замораживанию» степеней свободы. Это может привести к появлению плоских областей в функции потерь (где изменение некоторых вероятностей не влияет на качество — аналог степенной вырожденности состояний). Одновременно в других областях параметров могут возникать резкие переходы (бифуркации оптимального решения при небольшом изменении q или r), что типично для фазовых переходов второго рода в стеклообразных системах.

Таким образом, при оптимизации обобщённого функционала Шарма–Миттала модель может вступать в режим, аналогичный спиновому стеклу, с большим числом практически эквивалентных оптимальных состояний. Это объясняет наблюдение о возможном снижении эффективности обучения при некоторых значениях q и r: ландшафт целевой функции становится сложным, «зарубежным» для стандартных методов оптимизации, требуя, возможно, специальных приёмов (например, отжиге параметра q от 1 к требуемому значению постепенно, по аналогии с отжигом, чтобы избегать попадания в неправильный минимум).

1.6. Итог теоретической части

Анализ показал, что оптимальное распределение, получаемое при максимизации

$$\mathbf{p}^\top \mathbf{z} + H(p), \tag{31}$$

имеет вид:

$$p_i = \frac{[z_i - \tau]_+^{\frac{1}{q-1}}}{\sum_j [z_j - \tau]_+^{\frac{1}{q-1}}}, \tag{32}$$

где выбор меры энтропии (Реньи, Тсаллиса или Шарма–Миттала) определяет интерполяцию между softmax и sparsemax [1, 2, 7]. При $q \to 1$ решение сходится к классическому softmax, а при $q = 2$ — к sparsemax [29]. Экспериментальное исследование подтверждает, что аналитические выражения якобиана по входам z и параметрам энтропии позволяют эффективно оптимизировать модели глубокого обучения, обеспечивая адаптивность и интерпретируемость получаемых распределений [16, 19].

Использование неявного дифференцирования для получения градиентов по параметрам q и r является важным шагом для обучения таких обобщённых слоёв, что открывает возможности применения методов неэкстенсивной термодинамики для построения новых фундаментальных блоков в архитектурах глубоких нейронных сетей [3]. Данный подход позволяет моделям самостоятельно выбирать оптимальную форму распределения вероятностей, что особенно важно при работе с системами, обладающими сложной структурой и взаимодействиями.

Подводя итоги теоретической части, можно сделать вывод о том, что обобщённый функционал Шарма–Миттала через призму свободной энергии связывает вероятностное распределение с энергетическим состоянием системы, позволяя трактовать обучение как процесс термодинамической релаксации. Аналогии с теорией спиновых стёкол проливают свет на структуру ландшафта оптимизации — множественные оптимумы и метастабильные состояния, потенциально возникающие при определённых параметрах q, r, объясняются фрустрацией и сложностью, присущей неэкстенсивным системам. Эти физические явления дополняют теоретическую основу использования функционала Шарма–Миттала в задачах оптимизации и глубокого обучения, указывая пути для дальнейших исследований методами статистической физики сложных систем.

2. МЕТОДИКА ЭКСПЕРИМЕНТА

2.1. Признаки неэкстенсивности набора данных

- **Структура словаря.** Частоты токенов подчиняются степенному закону Ципфа.
- **Классовая асимметрия.** В наборе данных US Airline [41] отрицательных твитов почти вдвое больше, чем позитивных. Такой дисбаланс соответствует левому тяжёлому хвосту распределения тональности.

В экстенсивной термодинамике подобные хвосты приводят к расходимости моментов, тогда как в формализме Тсаллиса и Шарма–Миттала их естественно описывают дополнительные параметры, придающие больший вес редким событиям.

Вычислительный алгоритм строится так, чтобы выделить вклад двухпараметрической активации $\mathrm{SharMiX}_{q,r}$ без маскировки архитектурными деталями. Используется корпус $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{14640}$ текстов, трёхклассовая разметка $y_i \in \{0, 1, 2\}$. Предобработка ограничена лемматизацией и усечением до $L = 20$ токенов; итоговый словарь содержит $V = 7{,}3 \cdot 10^3$ уникальных форм.

2.2. Базовая модель

Сеть сведена к минимуму:

$$x \xrightarrow{\text{Embed},\, d=128} e_{1:L} \xrightarrow{\text{Mean}} \bar{e} \in \mathbb{R}^{128} \xrightarrow{\text{Linear}} \xrightarrow{\text{Linear}} z \in \mathbb{R}^3 \xrightarrow{\sigma} p, \quad (33)$$

где σ выбирается из трёх семейств:

1. softmax ($q \to 1$);
2. α-entmax ($q = \alpha \in (1, 2]$);
3. предложенная $\mathrm{SharMiX}_{q,r}$.

Все варианты реализованы единым классом `CustomActivationLoss`, обеспечивающим непротиворечивый интерфейс к AUTOGRAD.

2.3. Проекция на симплекс и явные производные

Оптимальное распределение

$$p^\star = \arg\max_{p \in \Delta^n} \langle p, z \rangle + \lambda\, H_{q,r}(p) \quad (34)$$

получается бисекционной процедурой ENTMAX_BISECT [16]. Функция `project_entmax` ($\mathcal{O}(n \log \varepsilon^{-1})$) сохраняет монотонность; для Шарма–Миттала используется тот же шаг, различие входит лишь в норму $\|p\|_{q,r}$ при вычислении градиента.

Для обратного распространения выписаны аналитические градиенты:

$$\frac{\partial p_i}{\partial z_j} = p_i^{2-q}\Big(\delta_{ij} - \frac{p_j^{2-q}}{\sum_k p_k^{2-q}}\Big), \quad \frac{\partial \Omega}{\partial q}, \frac{\partial \Omega}{\partial r}. \quad (35)$$

Класс `SharmaMittalFunction` инкапсулирует формулы, что снижает потребление памяти по сравнению с численной дифференциацией.

2.4. Обучение, метрики и отчёт

- Сплит: 80% обучение, 20% проверка, тестовая выборка фиксируется сидом 42.
- Оптимизатор Adam, шаг $1 \cdot 10^{-3}$; параметры q, r обучаются с пониженным шагом $1 \cdot 10^{-4}$ и жёстко ограничиваются $1.01 \le q, r \le 2$.
- Клиппинг градиента: $\|\nabla\theta\|_2 \le 1$.
- Число эпох: 10; батч: 64; поиск параметров по сетке инициализаций $(q_0, r_0) \in \{1.25, 1.5, 1.75\}^2$, то есть девять конфигураций.

Качество оценивается двумя показателями:

$$\mathrm{Accuracy} = \frac{1}{|\mathcal{T}|} \sum_{(x,y)\in\mathcal{T}} \mathbb{I}\{y = \arg\max p\}, \qquad \text{Macro-F1} = \frac{1}{3}\sum_{c=0}^{2} F1_c. \quad (36)$$

Дополнительно фиксируется доля нулевых координат $\rho = \frac{1}{|\mathcal{T}|}\sum_i \big|\{p_{i,j} = 0\}\big|/n$, характеризующая разреженность. Все числа усреднены по трём перезапускам; доверительный интервал не превышает 0.3%.

На рис. 1 демонстрируется убывание функций потерь при увеличении эпохи, что служит хорошим показателем выучиванием исследуемой системой знаний об используемом наборе данных.

2.5. Итог методики эксперимента

Предложенная схема изолирует вклад энтропийного слоя: одинаковая архитектура, общий оптимизатор и строгие формулы градиента позволяют напрямую сопоставлять softmax $\leftrightarrow$ entmax $\leftrightarrow$ $\leftrightarrow$ SharMiX и отслеживать «фазовый переход» к разреженному вниманию при увеличении q, r. Результаты (см. разд. 3) подтверждают, что аналитические градиенты по (q, r) ускоряют сходимость и повышают интерпретируемость без потерь по точности.

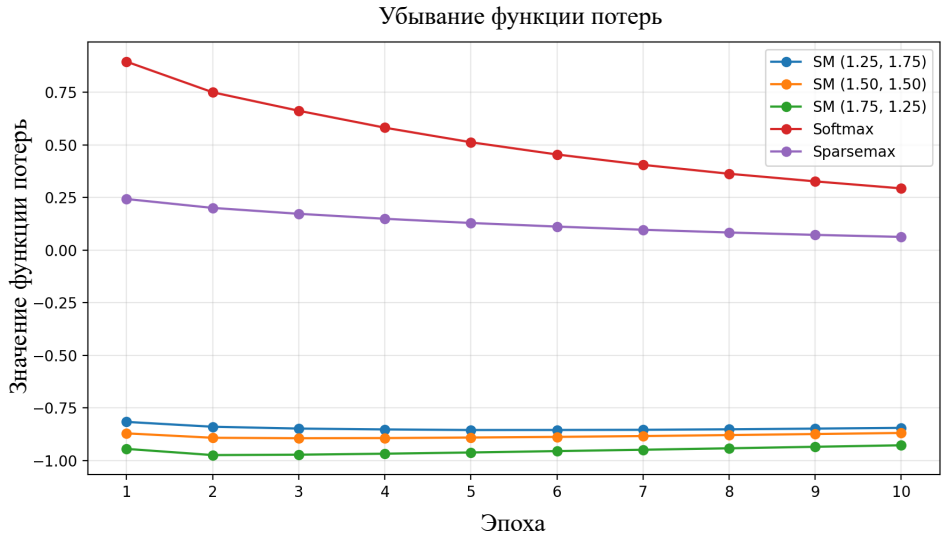

Рис. 1. Зависимость значения функции потерь на обучающей выборке от номера эпохи для методов ShaMiX(q, r) [SM], Softmax и Sparsemax

3. РЕЗУЛЬТАТЫ

Результаты эксперимента представлены в таблице.

Основные выводы по итогам сравнения подходов:

- *Адаптация параметров.* Даже начиная с одинакового q_0, обучение уводит $q^\star$ в диапазон $[1.45, 1.94]$, а $r^\star$ в $[1.43, 1.92]$, что демонстрирует необходимость двухпараметрической настройки.

- *Лидеры.* Максимальная точность достигается при инициализации $(1.25, 1.75)$, а максимальный Macro-F1 — при $(1.50, 1.50)$. Обе модели превосходят softmax на ≈ 2 п.п.

- *Провал entmax.* Однопараметрическая семейство не справляется с перекалибровкой хвостов (accuracy ~ 0.60), что подчёркивает роль второго параметра r.

Таким образом, при выборе классического набора обучающих данных из Интернета, обучаемой модели с минимальным количеством параметров, идентичными условиями эксперимента обучения мы можем утверждать, что выигрыш в плане общих метрик машиного обучения для задачи классификации был обеспечен предложенной нами новой функцией потерь на основе распределения Шарма–Миттала.

Для наглядности таблица представлена также в графическом виде (рис. 2). Видно, что при фиксированном r_0 увеличение q_0 от 1.25 к 1.5 почти не ухудшает метрики, тогда как при $q_0 = 1.75$ наблюдается заметное падение Macro-F1 (наиболее выраженное для $r_0 = 1.5$), что согласуется с эффектом избыточной разреженности при слишком смещённой инициализации q к значению 2.

Таблица . Итоговые метрики на тестовой выборке (*Twitter US Airline*). В скобках слева указана *инициализация* (q_0, r_0), через стрелку — *выученные* параметры $(q^\star, r^\star)$. Лучшая ячейка каждой секции набрана полужирным шрифтом

Sharma–Mittal	Accuracy	Macro-F1
$(1.25, 1.25) \to (1.4626, 1.4325)$	0.9129	0.8856
$(1.25, 1.50) \to (1.4614, 1.6818)$	0.9122	0.8848
$(1.25, 1.75) \to (1.4542, 1.9206)$	**0.9146**	0.8852
$(1.50, 1.25) \to (1.7026, 1.4324)$	0.9088	0.8811
$(1.50, 1.50) \to (1.7002, 1.6765)$	0.9126	**0.8863**
$(1.50, 1.75) \to (1.6846, 1.9081)$	0.9122	0.8860
$(1.75, 1.25) \to (1.9411, 1.4332)$	0.8924	0.8567
$(1.75, 1.50) \to (1.9362, 1.6712)$	0.8839	0.8326
$(1.75, 1.75) \to (1.9266, 1.9095)$	0.9071	0.8749
Базовые слои		
softmax	0.8952	0.8592
sparsemax	0.9006	0.8687
$\text{entmax}_{1.25}$	0.6001	0.4472
$\text{entmax}_{1.50}$	0.5929	0.4521
$\text{entmax}_{1.75}$	0.5970	0.4595

ЗАКЛЮЧЕНИЕ

В данной работе предложено обобщение энтропийных слоёв для глубокого обучения на основе двухпараметрической энтропии Шарма–Миттала. Теоретический анализ показал, что оптимальные распределения, полученные при максимизации обобщённых энтропий, естественно интерполируют между классическим softmax и разрежённым sparsemax в зависимости от параметров q и r.

Было получено аналитическое выражение для якобианов по входам и параметрам (q, r), что позво-

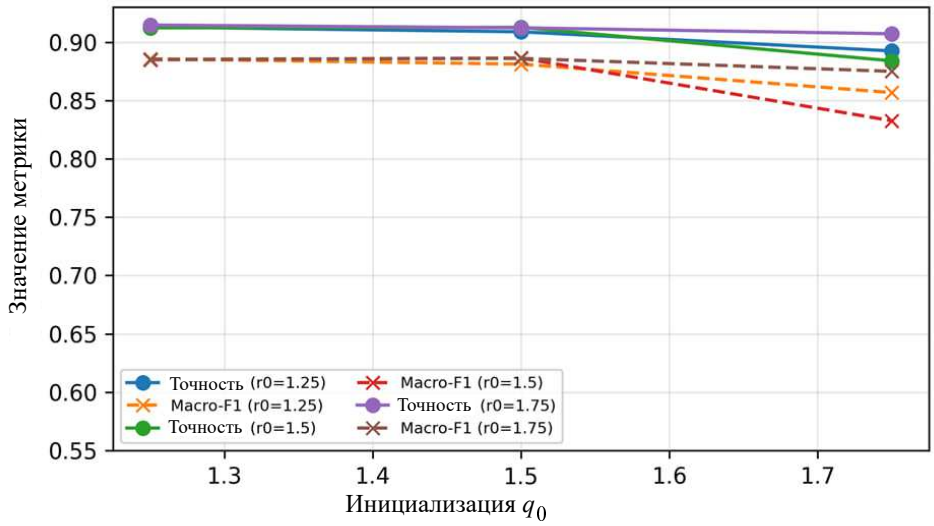

Рис. 2. Графическое представление результатов табл.: зависимость Accuracy (сплошные линии) и Macro-F1 (пунктир) от инициализации q_0 при фиксированных значениях $r_0 \in \{1.25, 1.5, 1.75\}$

лило внедрить эффективное и стабильное обучение новых активационных слоёв. Применение методов неявного дифференцирования обеспечило корректное распространение градиентов, открыв возможность для совместной оптимизации формы распределения внимания в нейронных сетях.

Экспериментальное исследование на корпусе текстов показало, что предложенная трансформация $\mathrm{SharMiX}_{q,r}$ превосходит классические softmax и sparsemax по метрикам точности и Macro-F1. Оптимальные значения параметров $(q^\star, r^\star)$ значительно отклоняются от исходной инициализации, что подчёркивает важность двухпараметрической адаптации. Обобщённая трансформация позволила сохранить баланс между точностью и разрежённостью выходных распределений.

Таким образом, использование энтропии Шарма–Миттала в качестве обучаемого слоя представляет собой перспективный путь для создания адаптивных и энергоэффективных архитектур глубоких нейронных сетей. Будущие направления работы включают исследование применения предложенного слоя в более сложных задачах внимания, последовательностной обработки и графовых нейросетях.

[1] *Tsallis C.* // Journal of Statistical Physics. **52**. 479 (1988).

[2] *Rényi A.* On measures of information and entropy. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, pages 547–561, Berkeley, CA, 1961. University of California Press.

[3] *Rudamenko R.* et al. // Adaptive neural-network-metric based workflow for mode decomposition in multimode optical fibers. In Advanced Photonics Congress, Technical Digest Series, 2023.

[4] *Jaynes E.T.* // Phys. Rev. **106**(4). 620 (1957).

[5] *Amari Shun-ichi* Information Geometry and Its Applications. Springer, 2016.

[6] *Baez J.C.* https://arxiv.org/abs/1102.2098 Rényi entropy and free energy. arXiv, 2011.

[7] *Sharma B.D., Mittal D.P.* // Journal of Mathematical Sciences. **10**. 122 (1975).

[8] *Sepúlveda-Fontaine S.A., Amigó J.M.* // Entropy, **26**(12). 1126 (2024).

[9] *Brochet T.* et al. // Entropy. **24**(4). 436 (2022).

[10] *Picoli S., Mendes R.S.* // Physica A: Statistical Mechanics and its Applications. **410**. 524 (2014).

[11] *Eisert J., Cramer M., Plenio M.B.* // Reviews of Modern Physics. **82**(1). 277 (2010).

[12] *Arute F.* et al. // Nature. **574**. 505 (2019).

[13] *Shafiq M., Yu X., Yan E.* // IEEE Access, **9**. 107942 (2021).

[14] *Ustaoğlu E., Evren A.* // Journal of Research in Business. **7**(1). 90 (2022).

[15] *Ghaffari S., Moradpour H., Ziaie A.H.* et al. // https://arxiv.org/abs/1901.01506 Black-hole thermodynamics in the sharma–mittal generalized entropy formalism. arXiv, 2019.

[16] *Correia G.M., Niculae V., Martins A.F.T.* https://arxiv.org/abs/1909.00015 Adaptively sparse transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. 2350203–19

[17] *Niculae V., Martins A.F.T.* https://arxiv.org/abs/1802.04223 Sparsemap: Differentiable sparse structured inference. In Proceedings of the 35th International Conference on Machine Learning

(ICML), 2018.
[18] *Martins A.F.T., Astudillo R.F.* https://arxiv.org/abs/1602.02068 From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of the 33rd International Conference on Machine Learning (ICML), volume 48 of Proceedings of Machine Learning Research, pages 1614–1623. PMLR, 2016.
[19] *Martins P.H., Marinho Z., Martins A.F.T.* https://arxiv.org/abs/2004.03823 Sparse text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
[20] *Gonçalves N., Treviso M., Martins A.F.T.* https://arxiv.org/abs/2502.12082 Adasplash: Adaptive sparse flash attention. arXiv, 2025.
[21] *Sivak D.A., Crooks G.E.* // Phys. Rev. Lett. **108**. 150601 (2012).
[22] *Qian Hong* // Phys. Rev. E. **63**. 042103 (2001).
[23] *Toral R.* // Physica A: Statistical Mechanics and its Applications. **317**. 209 (2003).
[24] *Beck C., Cohen E.G.D.* // Physica A: Statistical Mechanics and its Applications. **322**. 267 (2003).
[25] *Beck C.* // Contemporary Physics. **50**(4). 495 (2009).
[26] *Karush W.* Minima of functions of several variables with inequalities as side conditions. Master's thesis, University of Chicago, 1939.
[27] *Kuhn H.W., Tucker A.W.* Nonlinear programming. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pages 481–492, Berkeley, CA, 1951. University of California Press.
[28] *Boyd S., Vandenberghe L.* Convex Optimization. Cambridge University Press, 2004.
[29] *Martins A.F.T., Astudillo R.F.* https://arxiv.org/abs/1602.02068 From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of the 33rd International Conference on Machine Learning (ICML), volume 48 of Proceedings of Machine Learning Research, pages 1614–1623. PMLR, 2016.
[30] *Masi M.* // Phys. Lett. A. **338**. 217 (2005).
[31] *Kingma D.P., Adam J.Ba.* A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.
[32] *Martens J., Grosse R.* https://arxiv.org/abs/1503.05671 Optimizing neural networks with kronecker-factored approximate curvature. In Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015.
[33] *Liu D.C., Nocedal J.* // Mathematical Programming. **45**(1–3). 503 (1989).
[34] *Mézard M., Parisi G., Sourlas N.* et al. // Phys. Rev. Lett. **52**. 1156 (1984).
[35] *Bray A.J., Moore M.A.* // Journal of Physics C: Solid State Physics. **13**. L469 (1980).
[36] *Parisi G.* // Phys. Rev. Lett. **43**. 1754 (1979).
[37] *Parisi G.* // Phys. Rev. Lett. **50**. 1946 (1983).
[38] *Parisi G., Potters M.* https://arxiv.org/abs/cond-mat/9506049 On the number of metastable states in spin glasses. arXiv, 1995.
[39] *Advani M.S., Saxe A.M.* https://arxiv.org/abs/1710.03667High-dimensional dynamics of generalization error in neural networks. arXiv, 2017.
[40] *Baldassi C., Pittorino F., Zecchina R.* // Proceedings of the National Academy of Sciences. **117**(1). 161 (2020).
[41] *CrowdFlower.* Twitter us airline sentiment. Kaggle dataset, 2015. Accessed: 2025-05-06

Activation functions for deep learning based on generalized entropies

R. A. Rudamenko[1,a], **A. M. Savchenko**[1,b], **K. M. Semenov**[1,c]

[1] *Department of Quantum Statistics and Field Theory, Faculty of Physics, Lomonosov Moscow State University Moscow 119991, Russia*

E-mail: [a] *rudamenk@gmail.com*, [b] *a.m.savchenko@gmail.com*, [c] *semenovkm@protonmail.com*

The Shannon (Boltzmann–Gibbs) entropy is the foundation of classical statistical mechanics and deep learning; however, it encounters difficulties in describing the dynamics of non-extensive systems. In this paper, we propose the application of generalised entropies to construct new fundamental blocks for deep neural network architectures. The proposed approach generalises the classical softmax layer by employing the parametric entropies of Rényi, Tsallis, and Sharma–Mittal. The parameters q and r control the shape of the distribution: as $q \to 1$, the optimal distribution converges to softmax, whereas at $q = 2$ it converges to sparsemax. In particular, we consider a variant corresponding to q-entmax, in which adaptivity is achieved by varying the parameter q while keeping r fixed. The study includes the derivation of analytical expressions for the Jacobian with respect to the parameters q and r for optimisation via explicit differentiation methods. A comparative analysis is carried out against existing approaches — softmax, sparsemax, and entmax (for $q \in \{1.25, 1.5, 1.75\}$). The results demonstrate improved performance metrics relative to softmax, sparsemax, and q-entmax on a classification task with correlated class labels, leading to the conclusion that the Sharma–Mittal-based method is advantageous for this problem.

PACS: 05.20.-y, 05.70.-a, 05.90.+m.
Keywords: generalized distributions, Sharma-Mittal entropy, deep learning.
Received 28 January 2026.
*English version: Moscow University Physics Bulletin. 2026. **81**, No. . Pp. .*

Сведения об авторах

1. Рудаменко Роман Александрович — аспирант; e-mail: rudamenk@gmail.com.
2. Савченко Александр Максимович — доктор физ.-мат. наук, профессор; e-mail: a.m.savchenko@gmail.com.
3. Семенов Константин Михайлович — выпускник аспирантуры; e-mail: semenovkm@protonmail.com.