1 Introduction

The softmax function is widely used in game theory, reinforcement learning, and machine learning. Despite its popularity, some of its mathematical properties are not fully understood. This review focuses on these properties, based on the paper "On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning."

The paper uses convex analysis and monotone operator theory to reveal new properties of the softmax function. It shows that the softmax function is the gradient of the log-sum-exp function, which helps explain its behavior. The paper also highlights the importance of the inverse temperature parameter, , which affects the Lipschitz and co-coercivity properties of the softmax function.

In this review, we will:

Explain the mathematical properties of the softmax function in detail. Discuss how the parameter influences these properties.

2 Notations

The notations used in this paper are as follows:

- The p-norm of a vector is denoted as $\|\cdot\|_p$, $1 \leq p \leq \infty$.
- The $n-1$ dimensional unit simplex is denoted by Δ^{n-1}, where, $\Delta^{n-1} := \{x \in \mathbb{R}^n | \|x\|_1 = 1, x_i \geq 0\}$.
- The (relative) interior of Δ^{n-1} is denoted by $\text{int}(\Delta^{n-1})$, where, $\text{int}(\Delta^{n-1}) := \{x \in \mathbb{R}^n | \|x\|_1 = 1, x_i > 0\}$.
- $e_i \in \mathbb{R}^n$ denotes the i-th canonical basis of $\mathbb{R}^n$, e.g., $e_i = [0, \ldots, 1, \ldots, 0]^T$, where 1 occupies the i-th position.
- The vector of ones is denoted as $\mathbf{1} := [1, \ldots, 1]^T$ and the vector of zeros is denoted as $\mathbf{0} := [0, \ldots, 0]^T$.

Matrices are denoted using bold capital letters such as $\mathbf{A}$. In general, a vector in the unconstrained space $\mathbb{R}^n$ will be denoted using z, while a vector in the $n-1$ dimensional unit simplex will be denoted using x. All logarithms are assumed to be base e.

3 Review of the Softmax Function and its Known Properties

The softmax function can appear in various forms depending on the application. At its core, it is a vector-valued function where each component is an exponential of a vector element, normalized by the sum of these exponentials. In this section, we will show several common and equivalent forms of the softmax function and

review some basic properties that are immediately evident from its definition or are well-documented in the literature.

3.1 Representations of the Softmax function

The softmax function is given by $\sigma : \mathbb{R}^n \to \text{int}(\Delta^{n-1})$,
Definition 1:

$$\sigma(z) = \frac{1}{\sum_{j=1}^{n} \exp(\lambda z_j)} \begin{bmatrix} \exp(\lambda z_1) \\ \vdots \\ \exp(\lambda z_n) \end{bmatrix}, \quad \lambda > 0,$$

or

$$\sigma_i(z) = \frac{\exp(\lambda z_i)}{\sum_{j=1}^{n} \exp(\lambda z_j)}, \quad 1 \leq i \leq n.$$

$$\sigma_i(z) = \frac{\exp(\lambda z_i)}{\exp(\lambda z_i) + \exp(\lambda z_j)} = \frac{1}{1 + \exp(-\lambda(z_i - z_j))}, \quad j \neq i.$$

Furthermore, we note that (2) can be equivalently represented as,

$$\sigma_i(z) = \exp(\lambda z_i - \log(\sum_{j=1}^{n} \exp(\lambda z_j))).$$

Let $z \in \mathbb{R}^n$, and consider the arg max of $x^T z$ over the simplex,

$$M(z) := \arg \max_{x \in \Delta^{n-1}} x^T z.$$

However, it can produce multiple values when two or more components of z are the same. For many learning applications, it is very useful for $M(z)$ to produce a single value and also we define in simplex so:

$$M(z) = e_j, \quad \text{where} \quad j = \arg \max_{1 \leq i \leq n} e_i^T z.$$

A common method to ensure this is by using a regularizer function, which results in the regularized argmax function.

$$\tilde{M}(z) := \arg \max_{x \in \Delta^{n-1}} [x^T z - \psi(x)].$$

$$\psi(x) := \begin{cases} \lambda^{-1} \sum_{j=1}^{n} x_j \log(x_j), & \lambda > 0, x \in \Delta^{n-1} \\ +\infty, & x \notin \Delta^{n-1}. \end{cases}$$

Due to the strong concavity of the argument, applying the Karush-Kuhn-Tucker (KKT) conditions shows that the unique maximizer is the softmax function evaluated at $z \in \mathbb{R}^n$.

$$\arg \max_{x \in \Delta^{n-1}} \left[x^T z - \lambda^{-1} \sum_{j=1}^{n} x_j \log(x_j) \right] = \sigma(z).$$

$$\max_{x \in \Delta^{n-1}} \left[x^T z - \lambda^{-1} \sum_{j=1}^{n} x_j \log(x_j) \right] = lse(z).$$

$$\text{lse}(z) := \lambda^{-1} \log \left(\sum_{j=1}^{n} \exp(\lambda z_j) \right), \quad \lambda > 0.$$

$$\text{vecmax}(z) := \max\{z_1, \ldots, z_n\}.$$

$$\exp(\lambda \text{vecmax}(z)) \leq \sum_{j=1}^{n} \exp(\lambda z_j) \leq n \exp(\lambda \text{vecmax}(z)).$$

for any $z \in \mathbb{R}^n$, $\text{vecmax}(z) \leq \text{lse}(z) \leq \text{vecmax}(z) + \lambda^{-1} \log(n)$,
$\text{lse(z)} \geq x^T z - \psi(x), \forall x \in \Delta^{n-1}, z \in \mathbb{R}^n$.

Finally, we provide a probabilistic characterization of the softmax function. Let $\epsilon_i, i \in \{1, \ldots, n\}$ be independent and identically distributed random variables with a Gumbel distribution given by,

$$\Pr[\epsilon_i \leq c] = \exp(-\exp(-\lambda c - \gamma)),$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant. It can be shown that for any vector $z \in \mathbb{R}^n$

$$\Pr \left[i = \arg \max_{1 \leq j \leq n} \{z_j + \epsilon_j\} \right] = \sigma_i(z).$$

4 Properties of the Softmax - State of the Art

We briefly comment on some properties of the softmax function that are either immediate or have been covered in the existing literature. First, σ maps the origin of $\mathbb{R}^n$ to the barycenter of Δ^{n-1}, that is, $\sigma(0) = n^{-1}\mathbf{1}$. The softmax function σ is surjective but not injective, as it can easily be shown that for any $z, z + c\mathbf{1} \in \mathbb{R}^n, \forall c \in \mathbb{R}$, we have $\sigma(z + c\mathbf{1}) = \sigma(z)$.By definition,$\|\sigma(z)\|_1 = \sigma(z)^T \mathbf{1} = 1, \forall z \in \mathbb{R}^n$.

In a related direction, finding a lower bound on the softmax function. It can be demonstrated that, Firstly,notice that;

$$\frac{1}{a+b} \geq \frac{1}{1+a} \cdot \frac{1}{1+b}$$

Then case is multiple:

$$\sigma_i(z) = \frac{\exp(\lambda z_i)}{\sum_{j=1}^{n} \exp(\lambda z_j)} = \frac{1}{\sum_{j=1}^{n} \exp(-\lambda(z_i - z_j))} \geq \prod_{j=1, j \neq i}^{n} \frac{1}{1 + \exp(-\lambda(z_i - z_j))}$$

4.1 REVIEW OF CONVEX OPTIMIZATION AND MONOTONE OPERATOR THEORY

Definition 2: A function $f : \text{dom } f \subseteq \mathbb{R}^n \to \mathbb{R}$ is convex if,

$$f(\theta z + (1-\theta)z') \leq \theta f(z) + (1-\theta)f(z'), \quad \forall z, z' \in \text{dom } f \text{ and } \theta \in [0,1]$$

and strictly convex if (15) holds strictly whenever $z \neq z'$ and $\theta \in (0,1)$.

Lemma 1: Let f be C^2. Then f is convex if and only if dom f is convex and its Hessian is positive semidefinite, that is, for all $z \in \text{dom } f$, $v \in \mathbb{R}^n$,

$$v^T \nabla^2 f(z) v \geq 0,$$

and strictly convex if $\nabla^2 f(z)$ is positive definite for all $z \in \text{dom } f$.

Definition 3: An operator (or mapping) $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is said to be:

- pseudo monotone on D if,

$$F(z')^T(z-z') \geq 0 \implies F(z)^T(z-z') \geq 0, \quad \forall z, z' \in D.$$

- pseudo monotone plus on D if it is pseudo monotone on D and,

$$F(z')^T(z-z') \geq 0 \text{ and } F(z)^T(z-z') = 0 \implies F(z) = F(z'), \quad \forall z, z' \in D.$$

- monotone on D if,

$$(F(z) - F(z'))^T(z-z') \geq 0, \quad \forall z, z' \in D.$$

- monotone plus on D if it is monotone on D and,

$$(F(z) - F(z'))^T(z-z') = 0 \implies F(z) = F(z'), \quad \forall z, z' \in D.$$

- strictly monotone on D if,

$$(F(z) - F(z'))^T(z-z') > 0, \quad \forall z, z' \in D, z \neq z'.$$

Lemma 2: A C^1 function f is convex if and only if

$$(\nabla f(z) - \nabla f(z'))^T(z-z') \geq 0, \quad \forall z, z' \in \text{dom } f,$$

and strictly convex if and only if,

$$(\nabla f(z) - \nabla f(z'))^T(z-z') > 0, \quad \forall z, z' \in \text{dom } f, z \neq z'.$$

Definition 4: An operator (or mapping) $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is said to be

- Lipschitz (or L-Lipschitz) if there exists an $L > 0$ such that,

$$\|F(z) - F(z')\|_2 \leq L\|z - z'\|_2, \quad \forall z, z' \in D.$$

  If $L = 1$ in (24), then F is referred to as nonexpansive. Otherwise, if $L \in (0,1)$, then F is referred to as contractive.

- co-coercive (or $\frac{1}{L}$-co-coercive) if there exists an $L > 0$ such that,

$$(F(z) - F(z'))^T (z - z') \geq \frac{1}{L} \|F(z) - F(z')\|_2^2, \quad \forall z, z' \in D.$$

If $L = 1$ in (25), then F is referred to as firmly nonexpansive.

Baillon-Haddad theorem: Let $f : \text{dom } f \subseteq \mathbb{R}^n \to \mathbb{R}$ be a C^1, convex function on dom f and such that ∇f is L-Lipschitz continuous for some $L > 0$, then ∇f is $\frac{1}{L}$-co-coercive.

Definition 5: Let $H : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be monotone. Then H is maximal monotone if there exists no monotone operator $G : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ such that gra G properly contains gra H, i.e., for every $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$(u, v) \in \text{gra } H \iff (\forall (u', v') \in \text{gra } H)(u - u')^T (v - v') \geq 0.$$

If a continuous mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is monotone, it is maximal monotone. In particular, every differentiable monotone mapping is maximal monotone.

5 Derivation of Properties of Softmax Function

In this section, we use convex analysis and monotone operator theory to derive properties of the softmax function. We start by linking the log-sum-exp function to the softmax function. While it is known that the softmax function is the gradient of a convex potential function, it is less often noted that this potential function is the log-sum-exp function. This connection is clarified in the following proposition.

The softmax function is the gradient of the log-sum-exp function, that is,

$$\sigma(z) = \nabla \text{lse}(z)$$

.

Evaluating the partial derivative of lse at each component yields

$$\text{lse}(z) = \lambda^{-1} \log \left(\sum_{j=1}^{n} \exp(\lambda z_j) \right)$$

$$\frac{\partial \text{lse}(z)}{\partial z_i} = \frac{\exp(\lambda z_i)}{\sum_{j=1}^{n} \exp(\lambda z_j)} = \sigma_i(z)$$

By definition of the gradient, we have,

$$\nabla \text{lse}(z) = \begin{bmatrix} \frac{\partial \text{lse}(z)}{\partial z_1} \\ \vdots \\ \frac{\partial \text{lse}(z)}{\partial z_n} \end{bmatrix} = \frac{1}{\sum_{j=1}^{n} \exp(\lambda z_j)} \begin{bmatrix} \exp(\lambda z_1) \\ \vdots \\ \exp(\lambda z_n) \end{bmatrix} = \sigma(z).$$

Next, we calculate the Hessian of the log-sum-exp function (and hence the Jacobian of the softmax function).

The Jacobian of the softmax function and Hessian of the log-sum-exp function is given by:

$$J[\sigma(z)]_{ij} = \frac{\partial \sigma_i(z)}{\partial z_j} = \frac{\partial^2 \text{lse}(z)}{\partial z_j \partial z_i}$$

Propostion 1: $J[\sigma(z)]$ is a symmetric positive semidefinite matrix and satisfies $J[\sigma(z)]\mathbf{1} = 0$, that is, $\mathbf{1}$ is the eigenvector associated with the zero eigenvalue of $J[\sigma(z)]$ and also:

$$J[\sigma(z)] = H(\text{lse}(z)) = \nabla^2 \text{lse}(z) = \lambda(\text{diag}(\sigma(z)) - \sigma(z)\sigma(z)^T),$$

$$\text{diag}(\sigma(z))_{ij} := \begin{cases} \sigma_i(z) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

The diagonal entries of $\nabla^2 \text{lse}$ are given by,

$$\frac{\partial^2 \text{lse}(z)}{\partial z_i^2} = \lambda \left[\frac{\exp(\lambda z_i)}{\sum_{j=1}^n \exp(\lambda z_j)} - \frac{\exp(\lambda z_i)^2}{(\sum_{j=1}^n \exp(\lambda z_j))^2} \right],$$

and the off-diagonal entries of $\nabla^2 \text{lse}$ are given by the mixed partials,

$$\frac{\partial^2 \text{lse}(z)}{\partial z_k \partial z_i} = -\lambda \frac{\exp(\lambda z_k) \exp(\lambda z_i)}{(\sum_{j=1}^n \exp(\lambda z_j))^2}.$$

Proof: Remember that:

$$H(f(x)) = J(\nabla f(x))^T$$

$$H(\text{lse}(z))_{ij} = \frac{\partial^2 \text{lse}(z)}{\partial z_j \partial z_i}$$

For positive semidefinite:

$$x^T(\text{diag}(\sigma(z)) - \sigma(z)\sigma(z)^T)x = x^T \text{diag}(\sigma(z))x - x^T\sigma(z)\sigma(z)^T x = \sum_{i=1}^n x_i^2 \sigma_i(z) - \sum_{i=1}^n x_i \sigma_i(z) \sum_{j=1}^n x_j \sigma_j(z) \geq 0$$

by Cauchy–Schwarz inequality
For 0 eigenvalue:

$$J[\sigma(z)]\mathbf{1} = \sigma(z) - \sigma(z) \left(\sum_{i=1}^n \sigma_i(z) \right) = \sigma(z) \left[1 - \sum_{i=1}^n \sigma_i(z) \right] = 0$$

Lemma 4: The log-sum-exp function is C^2, convex, and not strictly convex on $\mathbb{R}^n$.ie:

$$x^T \nabla^2 \text{lse}(z) x \geq 0$$

Proof: It is straight forward from lemma 1 and previos proposition 1.
Proposition 3: The softmax function is monotone, that is,

$$(\sigma(z) - \sigma(z'))^T (z - z') \geq 0, \quad \forall z, z' \in \mathbb{R}^n;$$

and not strictly monotone on $\mathbb{R}^n$.
Proof:

$$(\sigma(z) - \sigma(z'))^T (z - z') = (\nabla \text{lse}(z) - \nabla \text{lse}(z'))^T (z - z') \geq 0, \quad \forall z, z' \in \mathbb{R}^n;$$

, $\forall z, z' \in \mathbb{R}^n$; and wit lemma 2 proof is complete.
Corollary 1: The softmax function is a maximal monotone operator, that is, there exists no monotone operator such that its graph properly contains the graph of the softmax function.
Proof: This directly follows from being a continuous, monotone map from Lemma 3. Next, we show that under appropriate conditions, the softmax function is a contraction in $\|\cdot\|_2$.

Lemma 5: A C^2, convex function $\text{lse} : \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz continuous gradient with Lipschitz constant $L > 0$ if for all $z, x \in \mathbb{R}^n$,

$$0 \leq x^T \nabla^2 \text{lse}(z) x \leq L\|x\|_2^2.$$

Proposition 4: The softmax function is L-Lipschitz with respect to $\|\cdot\|_2$ with $L = \lambda$, that is, for all $z, z' \in \mathbb{R}^n$,

$$\|\sigma(z) - \sigma(z')\|_2 \leq \lambda \|z - z'\|_2,$$

where λ is the inverse temperature constant.

The softmax function is L-Lipschitz with respect to $\|\cdot\|_2$ with $L = \lambda$, that is, for all $z, z' \in \mathbb{R}^n$,

$$\|\sigma(z) - \sigma(z')\|_2 \leq \lambda \|z - z'\|_2,$$

where λ is the inverse temperature constant.

Proof. Given the Hessian of lse in Proposition 2, we have for all $z, x \in \mathbb{R}^n$,

$$x^T \nabla^2 \text{lse}(z) x = \lambda \left(\sum_{i=1}^n x_i^2 \sigma_i(z) - \left(\sum_{i=1}^n x_i \sigma_i(z) \right)^2 \right).$$

Since the second term on the right hand side of above is nonnegative, therefore,

$$x^T \nabla^2 \text{lse}(z) x \leq \lambda \sum_{i=1}^n x_i^2 \sigma_i(z) \leq \lambda \sup\{\sigma_i(z)\} \sum_{i=1}^n x_i^2 \Rightarrow x^T \nabla^2 \text{lse}(z) x \leq \lambda \|x\|_2^2.$$

where $\sup\{\sigma_i(z)\} = 1$, $\forall i \in \{1, \ldots, n\}$, $\forall z \in \mathbb{R}^n$. By our Lemma , $\nabla^2 \text{lse}(z)$ is positive semidefinite. Hence, we have,

$$0 \leq x^T \nabla^2 \text{lse}(z) x \leq \lambda \|x\|_2^2.$$

By Lemma 5, σ is Lipschitz with $L = \lambda$.

As a minor consequence of Proposition 4, by the Cauchy-Schwarz inequality, we have,

$$(\sigma(z) - \sigma(z'))^T(z - z') \leq \lambda \|z - z'\|_2^2.$$

Corollary 2: The softmax function is $\frac{1}{L}$-co-coercive with respect to $\|\cdot\|_2$ with $L = \lambda$, that is, for all $z, z' \in \mathbb{R}^n$,

$$(\sigma(z) - \sigma(z'))^T(z - z') \geq \frac{1}{\lambda}\|\sigma(z) - \sigma(z')\|_2^2,$$

where λ is the inverse temperature constant.

Proof. Follows directly from Baillon-Haddad Theorem.

6 Conclusion

The Softmax function whose value is the index of a vector's largest element.The softmax function serves as a smooth approximation to the arg max function, which identifies the index of the largest element in a vector.This paper thoroughly analyzes the softmax function using convex analysis and monotone operator theory. We demonstrate that the softmax function is the monotone gradient map of the log-sum-exp function and that the inverse temperature parameter λ determines its Lipschitz and co-coercivity properties. These properties help construct convergence guarantees for score dynamics in various games . The reinforcement learning scheme structure is similar to those in bandit and online learning, such as the FTRL and mirror descent algorithms .