Efficient First-Order Algorithms for Large-Scale, Non-Smooth Maximum Entropy Models with Application to Wildfire Science

Gabriel Provencher Langlois [1,*], **Jatan Buch** [2] and **Jérôme Darbon** [3]

1   Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
2   Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA; jb4625@columbia.edu
3   Division of Applied Mathematics, Brown University, Providence, RI 02912, USA; jerome_darbon@brown.edu
*   Correspondence: gabriel.provencher@courant.edu

Abstract: Maximum entropy (MaxEnt) models are a class of statistical models that use the maximum entropy principle to estimate probability distributions from data. Due to the size of modern data sets, MaxEnt models need efficient optimization algorithms to scale well for big data applications. State-of-the-art algorithms for MaxEnt models, however, were not originally designed to handle big data sets; these algorithms either rely on technical devices that may yield unreliable numerical results, scale poorly, or require smoothness assumptions that many practical MaxEnt models lack. In this paper, we present novel optimization algorithms that overcome the shortcomings of state-of-the-art algorithms for training large-scale, non-smooth MaxEnt models. Our proposed first-order algorithms leverage the Kullback–Leibler divergence to train large-scale and non-smooth MaxEnt models efficiently. For MaxEnt models with discrete probability distribution of n elements built from samples, each containing m features, the stepsize parameter estimation and iterations in our algorithms scale on the order of $O(mn)$ operations and can be trivially parallelized. Moreover, the strong ℓ_1 convexity of the Kullback–Leibler divergence allows for larger stepsize parameters, thereby speeding up the convergence rate of our algorithms. To illustrate the efficiency of our novel algorithms, we consider the problem of estimating probabilities of fire occurrences as a function of ecological features in the Western US MTBS-Interagency wildfire data set. Our numerical results show that our algorithms outperform the state of the art by one order of magnitude and yield results that agree with physical models of wildfire occurrence and previous statistical analyses of wildfire drivers.

Keywords: primal–dual method; maximum entropy estimation; Kullback–Leibler divergence; wildfire science

MSC: 90C30; 90C06; 90C90; 62P12

1. Introduction

Maximum entropy (MaxEnt) models are a class of density estimation methods that use the maximum entropy principle [1] to reproduce key statistics of data sets. Historically used in physics [1,2], engineering [3–5], and statistics [6] applications, MaxEnt models are now frequently used for large-scale machine learning problems in natural language processing [7–12], social science [13–15], neuroscience [16–18], ecological modeling [19–28], and wildfire science [29–31]. MaxEnt models in machine learning must often estimate probability distributions from big data sets comprising hundreds of thousands to billions of samples or features or both [32]. Due to this enormous number, large-scale MaxEnt models need efficient and robust algorithms to perform well.

State-of-the-art algorithms for MaxEnt models, however, were not originally designed to handle big data sets; these algorithms either rely on technical devices that may yield unreliable numerical results [33], scale poorly in size [9,34], or require smoothness assumptions that many MaxEnt models lack in practice [10]. These limitations make it essentially

impossible to scale MaxEnt models to big data applications without adequate and costly computational resources [35,36]. This constraint on computational resources, in particular, has been recently identified as a crucial challenge to overcome for using large-scale MaxEnt models in climate change and wildfire studies within a reasonable amount of run time [27,28,30,37].

1.1. Contributions of This Paper

In this paper, we present novel optimization algorithms that overcome the shortcomings of state-of-the-art algorithms used for training large-scale, non-smooth MaxEnt models. Our proposed algorithms are first-order accelerated nonlinear primal–dual hybrid gradient (NPDHG) algorithms, whose theory was provided by two authors of this paper in [38], based on the Kullback–Leibler divergence. Using the Kullback–Leibler divergence over the classical proximal operator makes it possible to train large-scale and non-smooth MaxEnt models much more efficiently than the state of the art. In particular, we show that, for a MaxEnt model with a discrete probability distribution of n elements built from samples each containing m features, the stepsize parameter estimation and iterations in our algorithms all scale on the order of $O(mn)$ operations. This significantly improves on the known complexity bound of $O(\min(m^2n, mn^2))$ operations for computing the optimal stepsize parameters of classical first-order optimization methods, such as the linear PDHG or forward–backward splitting methods. We also show, as a consequence, that, for a given tolerance level $\epsilon > 0$, our algorithms provably compute solutions using on the order of $O(mn/\sqrt{\epsilon})$ or $O(mn/\log(1/\epsilon))$ operations, the order depending on the smoothness of the MaxEnt model and which are optimal with respect to the Nesterov class of optimal first-order methods [39]. Moreover, the computational bottleneck consists of matrix–vector multiplications, which can be trivially parallelized, and so our proposed algorithms exhibit scalable parallelism. Finally, we show that the strong convexity of the Kullback–Leibler divergence with respect to the ℓ_1 norm naturally allows for significantly larger stepsize parameters, thereby substantially speeding up the convergence rate of our algorithms.

To illustrate the efficiency of our novel algorithms on large-scale problems, we present an application to wildfire science. Specifically, we consider the problem of estimating the probabilities of fire occurrences across the western US as a function of ecological features. To do so, we fit elastic net, non-overlapping group lasso, and ℓ_∞ MaxEnt models to a large number of hyperparameters using our proposed algorithms and the state-of-the-art forward–backward splitting and coordinate descent STRUCTMAXENT2 algorithms [40,41]. Our numerical results show that our algorithms outperform the two latter algorithms by at least one order of magnitude, and yield results that are in good agreement with physical models of wildfire occurrence [42,43] as well as previous statistical analyses of wildfire drivers [29–31,44].

1.2. Organization of This Paper

This paper is organized as follows. In Section 2, we present our setup, we describe how MaxEnt models work, and we review three popular non-smooth MaxEnt models that are challenging to train with big data sets: the elastic net, group lasso, and ℓ_∞ regularized MaxEnt models. In Section 3, we explain why large-scale, non-smooth MaxEnt models are particularly challenging to train from big data sets, and we describe the state-of-the-art algorithms for training these MaxEnt models and their limitations. In Section 4, we describe our approach for estimating MaxEnt models with NPDHG optimization methods, derive explicit algorithms for regularized MaxEnt models, and explain why our algorithms overcome the limitations of state-of-the-art algorithms. In Section 5, we present an application of our algorithms to wildfire science, where we train a large number of MaxEnt models on the Western US MTBS-Interagency (WUMI) wildfire data set [45] to estimate the probabilities of fire occurrences as a function of ecological features. Finally, we review our results and outline directions for future work in Section 6.

2. Preliminaries

2.1. Setup

Suppose we receive l independent, identically distributed samples $\{v_1, \ldots, v_l\} \subset \mathcal{I}$ from an unknown distribution $\mathcal{D}$. We assume throughout, without of loss of generality, that the input space $\mathcal{I} \equiv \{1, \ldots, n\}$. In addition, suppose we have a prior probability distribution p_{prior} on $\mathcal{I}$ that encapsulates some prior knowledge about the samples or unknown distribution. Finally, suppose we have access to a set of features from the samples via a bounded feature map $\boldsymbol{\Phi} \colon \mathcal{I} \to \mathbb{R}^m$. Then, how do we estimate the unknown distribution $\mathcal{D}$ from the prior distribution p_{prior}, the samples $\{v_1, \ldots, v_l\}$, and the feature map $\boldsymbol{\Phi}$?

The maximum entropy principle offers a systematic way to answer this question. It states that the distribution that best estimates the unknown distribution $\mathcal{D}$ is the one that remains as close as possible to the prior probability p_{prior} while matching the features $\{\boldsymbol{\Phi}(v_1), \ldots, \boldsymbol{\Phi}(v_l)\}$ exactly or as closely as possible, in some suitable sense. Specifically, let Δ_n denote the n-dimensional probability simplex and suppose $p_{\text{prior}} \in \text{int}\,\Delta_n$ (i.e., $p_{\text{prior}}(j) > 0$ for every $j \in \{1, \ldots, n\}$). We measure the closeness of a probability distribution $p \in \Delta_n$ to the prior probability p_{prior} using the Kullback–Leibler divergence $D_{\text{KL}} \colon \Delta_n \times \text{int}\,\Delta_n \to \mathbb{R}$:

$$D_{\text{KL}}(p \parallel p_{\text{prior}}) = \sum_{j=1}^{n} p(j) \log\big(p(j)/p_{\text{prior}}(j)\big). \tag{1}$$

Next, let $\hat{\mathcal{D}}$ denote the empirical distribution induced by samples $\{v_1, \ldots, v_l\}$:

$$\hat{\mathcal{D}}(j) = \frac{1}{l} |\{1 \leqslant i \leqslant l \mid v_i = j\}|. \tag{2}$$

In addition, let $\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]$ denote the empirical average of the features induced by the samples and let $\mathbb{E}_p[\boldsymbol{\Phi}]$ denote the model average induced by a probability distribution $p \in \Delta_n$:

$$\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] = \sum_{j=1}^{n} \hat{\mathcal{D}}(j)\boldsymbol{\Phi}(j) \quad \text{and} \quad \mathbb{E}_p[\boldsymbol{\Phi}] = \sum_{j=1}^{n} p(j)\boldsymbol{\Phi}(j).$$

We measure how the averages $\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]$ and $\mathbb{E}_p[\boldsymbol{\Phi}]$ match using a function $H^* \colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ called the potential function. MaxEnt models combine the two measures described above and seek to minimize the sum

$$\inf_{p \in \Delta_n} f(p; t) = \inf_{p \in \Delta_n} \left\{ D_{\text{KL}}(p \parallel p_{\text{prior}}) + tH^*\left(\frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}]}{t}\right) \right\} \tag{3}$$

where $t > 0$ is a hyperparameter selected using data-driven methods, e.g., cross-validation.

Standing assumptions: The potential function is proper, lower semicontinuous, convex, and bounded from below by zero with $H^*(\mathbf{0}) = 0$. Under the standing assumptions above on the potential function, which are fairly general, the probability distribution $\hat{\mathcal{D}}$ is a feasible point of (3). Moreover, the MaxEnt estimation problem (3) admits a unique global solution (see ([46], Page 35, Proposition 1.2); uniqueness follows from the strong convexity of the Kullback–Leibler divergence with respect to the ℓ_1 norm [47–51]).

2.2. Dual Formulation and Optimality Conditions

The generalized MaxEnt problem (3) admits a dual problem corresponding to regularized maximum a posteriori estimation [52]. To derive the dual problem, we first write the second term on the right hand side of (3) in terms of its convex conjugate,

$$tH^*\left(\frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}]}{t}\right) = \sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}] \rangle - tH(\boldsymbol{w}) \right\},$$

where we abuse notation and write the convex conjugate of H^* as H. This lets us express problem (3) in saddle-point form:

$$\inf_{p \in \Delta_n} \sup_{w \in \mathbb{R}^m} \left\{ \langle w, \mathbb{E}_{\hat{\mathcal{D}}}[\Phi] - \mathbb{E}_p[\Phi] \rangle - tH(w) + D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) \right\}. \tag{4}$$

Due to our assumptions on the potential function H^*, we can swap the infimum and supremum ([46], Statement (4.1) on page 61). Using the convex conjugate formula

$$\inf_{p \in \Delta_n} \left\{ D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) - \langle w, \mathbb{E}_p[\Phi] \rangle \right\} = -\log \left[\sum_{j=1}^n p_{\mathrm{prior}}(j) e^{\langle w, \Phi(j) \rangle} \right],$$

we obtain the dual problem of (3)

$$\sup_{w \in \mathbb{R}^m} \left\{ \langle w, \mathbb{E}_{\hat{\mathcal{D}}}[\Phi] \rangle - tH(w) - \log \left[\sum_{j=1}^n p_{\mathrm{prior}}(j) e^{\langle w, \Phi(j) \rangle} \right] \right\}. \tag{5}$$

The dual problem (5) is a regularized maximum likelihood estimation problem over the family of Gibbs distributions [41,52]. It has at least one global solution ([46], Proposition (1.2) on page 35). Moreover, if (p^s, w^s) denotes a pair of solutions to the MaxEnt estimation problem (3) and its dual (5), then this pair satisfies the optimality conditions

$$\mathbb{E}_{\hat{\mathcal{D}}}[\Phi] - \mathbb{E}_{p^s}[\Phi] \in t\partial H(w^s) \quad \text{and} \quad p^s(j) = \frac{p_{\mathrm{prior}}(j) e^{\langle w^s, \Phi(j) \rangle}}{\sum_{j=1}^n p_{\mathrm{prior}}(j) e^{\langle w^s, \Phi(j) \rangle}}, \tag{6}$$

for every $j \in \{1, \dots, n\}$ where $\partial H(w^s)$ denotes the subdifferential of H at w^s.

2.3. Examples of Non-Smooth MaxEnt Models

MaxEnt models differ in the choice of the prior distribution p_{prior}, hyperparameter t, and the potential function H^*. In practice, the prior distribution is often chosen to be uniform and the hyperparameter is chosen via data-driven methods. The choice of potential function depends on the application. The classical MaxEnt model uses the indicator function

$$u \mapsto H^*(u) = \begin{cases} 0, & \text{if } u = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

This forces the model average $\mathbb{E}_p[\Phi]$ to be equal to the empirical average $\mathbb{E}_{\hat{\mathcal{D}}}[\Phi]$. However, this is often too restrictive because the averages are expected to be close, and not equal, with high probability. Forcing the model and empirical averages to be equal can lead to severe over-fitting issues with big data sets, and more flexibility is therefore desired.

We review below three different non-smooth MaxEnt models that offer this flexibility: the elastic net, group lasso, and ℓ_∞-regularized MaxEnt models. We chose these models because they are used extensively in statistics and machine learning but are also challenging to train on big data sets due to their non-smoothness [20,26,27,53]. These models will also be used for our numerical experiments and results in Section 5.

2.3.1. Elastic Net Regularized MaxEnt Models

These models use for potential function the relaxation of the convex conjugate of the ℓ_1-norm:

$$u \mapsto H^*(u) \equiv \left(\frac{1-\alpha}{2} \|\cdot\|_2^2 + \alpha \|\cdot\|_1 \right)^* (u) = \frac{1}{2(1-\alpha)} \sum_{i=1}^m \max(0, |u_i| - \alpha)^2$$

where $\alpha \in (0, 1)$. This yields the elastic net regularized maximum entropy estimation problem

$$\min_{p \in \Delta_n} \left\{ D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) + \frac{1}{2t(1-\alpha)} \sum_{i=1}^{m} \max\left(0, \left| [\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}]]_i - [\mathbb{E}_p[\boldsymbol{\Phi}]]_i \right| - t\alpha \right)^2 \right\}. \tag{7}$$

This type of regularization is frequently used for feature selection in machine learning [12,20,54,55]. The corresponding dual problem is

$$\sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] \rangle - t \left(\frac{(1-\alpha)}{2} \|\boldsymbol{w}\|_2^2 + \alpha \|\boldsymbol{w}\|_1 \right) - \log \left[\sum_{j=1}^{n} p_{\mathrm{prior}}(j) e^{\langle \boldsymbol{w}, \boldsymbol{\Phi}(j) \rangle} \right] \right\}. \tag{8}$$

Thanks to the elastic net penalty $\frac{(1-\alpha)}{2} \|\boldsymbol{w}\|_2^2 + \alpha \|\boldsymbol{w}\|_1$, problem (8) is strongly concave and has a unique solution.

An important aspect of the elastic net penalty is that it promotes sparsity; that is, it leads to a solution with many entries equal to zero [56–58], the number depending on the hyperparameter t and parameter α. Sparsity is useful when seeking accurate solutions based only on a few features or when performing feature selection.

2.3.2. Group Lasso Regularized MaxEnt Models

These models are specified starting from the dual problem. Consider a partition G of the components of a vector $\boldsymbol{w} \in \mathbb{R}^m$ into g different and possibly overlapping groups $[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_g]$ with $\boldsymbol{w}_g \in \mathbb{R}^{m_g}$ and $\cup_{g=1}^{G} \boldsymbol{w}_g = \boldsymbol{w}$. Let $\boldsymbol{w} \mapsto \|\boldsymbol{w}_g\|_{2,g}$ denote the ℓ_2 norm over the respective components of the gth group. Then, the dual version of the group lasso regularized maximum entropy estimation problem is

$$\sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] \rangle - t \sum_{g=1}^{G} \sqrt{m_g} \|\boldsymbol{w}_g\|_{2,g} - \log \left[\sum_{j=1}^{n} p_{\mathrm{prior}}(j) e^{\langle \boldsymbol{w}, \boldsymbol{\Phi}(j) \rangle} \right] \right\}. \tag{9}$$

Thanks to the group lasso penalty $\sum_{g=1}^{G} \sqrt{m_g} \|\boldsymbol{w}_g\|_{2,g}$, problem (10) is strictly concave and so has at least one global solution. The corresponding primal problem follows from the convex conjugate formula

$$\boldsymbol{u} \mapsto H^*(\boldsymbol{u}) \equiv \left(\sum_{g=1}^{G} \sqrt{m_g} \|\cdot\|_{2,g} \right)^* (\boldsymbol{u}) = \begin{cases} 0, & \text{if } \|\boldsymbol{u}_g\|_{2,g} \leqslant \sqrt{m_g} \text{ for } g = 1, \ldots, G, \\ +\infty, & \text{otherwise}, \end{cases}$$

and it is therefore given by

$$\min_{p \in \Delta_n} D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) \quad \text{s.t.} \quad \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}_g] - \mathbb{E}_p[\boldsymbol{\Phi}_g] \right\|_{2,g} \leqslant t\sqrt{m_g} \quad \text{for} \quad g = 1, \ldots, G. \tag{10}$$

Similarly to the elastic net penalty, the group lasso penalty promotes sparsity in the global solutions of (9), but it differs in that it sets different blocks of components $[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_g]$ to zero at the same time.

2.3.3. ℓ_∞-Regularized MaxEnt Models

These models use for potential function the convex conjugate of the ℓ_∞ norm,

$$\boldsymbol{u} \mapsto H^*(\boldsymbol{u}) \equiv \|\boldsymbol{u}\|_\infty^* = \{\hat{\boldsymbol{u}} \in \mathbb{R}^m \mid \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_1 \leqslant 1\},$$

which is the characteristic set of the ℓ_1 ball with unit radius. This yields the ℓ_∞-regularized maximum entropy estimation problem

$$\min_{p \in \Delta_n} \left\{ D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) + t \left\| \frac{\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_p[\boldsymbol{\Phi}]}{t} \right\|_\infty^* \right\}. \tag{11}$$

This type of regularization is used for certain machine learning problems, including matrix factor decomposition, and in robust statistics applications (see, e.g., [53,59]). The corresponding dual problem is

$$\sup_{\boldsymbol{w} \in \mathbb{R}^m} \left\{ \langle \boldsymbol{w}, \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] \rangle - t\|\boldsymbol{w}\|_\infty - \log\left[\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle \boldsymbol{w}, \boldsymbol{\Phi}(j) \rangle} \right] \right\} \tag{12}$$

Thanks to the ℓ_∞ norm, problem (12) is strictly concave and has at least one global solution.

3. Related Work

3.1. Large-Scale Sparse MaxEnt Models: Computational Challenges

Estimating a probability distribution from the MaxEnt model (3) can be computationally prohibitive for big data sets. To illustrate this point, suppose the MaxEnt model (3) with hyperparameter $t \geqslant 0$ has a global solution $p^s(t)$. Let $p^\epsilon(t) \in \Delta_n$ with $\epsilon > 0$ denote an ϵ-approximate solution to the global solution $p^s(t)$, that is, the objective function f in (3) satisfies

$$f(p^\epsilon(t); t) - f(p^s(t); t) < \epsilon.$$

If the potential function H^* is smooth (equivalently, its conjugate H is strongly convex), the best achievable rate of convergence for computing $p^\epsilon(t)$ with $t > 0$ in the Nesterov class of optimal first-order methods is linear $O(\log(1/\epsilon))$ in the number of operations [39]. If H is not smooth, then the optimal convergence rate is sublinear $O(1/\sqrt{\epsilon})$ in the number of operations.

These rates, while optimal, require carefully fine-tuned stepsize parameters. In classical first-order optimization algorithms, these stepsize parameters are fine-tuned using a precise estimate of the largest singular value of the linear operator $A \colon \Delta_n \to \mathbb{R}^m$ defined by

$$Ap = \sum_{j=1}^n p(j)\boldsymbol{\Phi}(j) = \mathbb{E}_p[\boldsymbol{\Phi}]. \tag{13}$$

Unfortunately, computing the largest singular value of the linear operator A in (13) accurately is often computationally expensive for large matrices due to its prohibitive cubic cost of $O(\min(m^2 n, mn^2))$ operations [60]. In this situation, line search methods and other heuristics can sometime be employed to bypass this issue, but they typically slow down the convergence speed. Another approach is to compute a crude estimate of the largest singular value of the matrix A, but doing so significantly reduces convergence speed as well. In fact, even if the largest singular value can be computed quickly, the resulting stepsize parameters may be much smaller than what is permissible to maintain convergence, that is, the largest singular value itself may be a poor estimate for determining how large the stepsize parameters are allowed to be to maintain convergence. This point will discussed in more detail in Section 4.1 when we describe our proposed NPDHG algorithms.

This issue makes solving the MaxEnt model (3) difficult and laborious. Even worse, in some applications, the appropriate value of the hyperparameter t in (3) is difficult to guess and must be selected by repeatedly solving (3) from a large pool of values of hyperparameters, a process that can become particularly time-consuming and resource intensive for big data sets. This issue has driven much research in developing robust and efficient algorithms to minimize computational costs and maximize model performance.

3.2. State-of-the-Art Methods for Large-Scale, Non-Smooth MaxEnt Models

State-of-the-art methods for computing solutions to large-scale, non-smooth MaxEnt models are based on coordinate descent algorithms [61–64] and first-order optimization algorithms such as forward–backward splitting [40,65,66]. We will discuss these methods below, but note that other types of methods have been developed for MaxEnt models; see, e.g., [10,41,67,68] for surveys and comparisons of different algorithms. In particular,

we will not consider second-order-based methods suitable for smooth MaxEnt models such as limited-memory BFGS algorithms [10,69] since this work focuses on non-smooth MaxEnt models.

Coordinate descent methods. The state of the art is a coordinate descent algorithm based on a technical device called infinitely weighted logistic regression (IWLR) [26,33]. The IWLR method approximates a MaxEnt model as a logistic regression model and then fits the approximate logistic regression model using an existing, efficient optimization algorithm for logistic regression. This indirect approach was proposed because efficient and scalable coordinate descent algorithms were already available for fitting logistic regression models, which in [33] was an earlier version of the GLMNET software package [63]. The IWLR method remains the state of the art for this reason. It is implemented, for example, in the MaxEnt package available in the R programming language [26,63].

The IWLR method, however, is an approximate technical device that is not guaranteed to work, a fact acknowledged by the authors who proposed the method [33], and therefore may not produce reliable numerical results. Coordinate descent algorithms themselves are generally non-parallelizable and often lack robustness and good convergence properties. For example, the aforementioned GLMNET software package approximates the logarithm term in logistic regression models with a quadratic term to fit the models efficiently. Without costly stepsize optimization, which GLMNET avoids to improve performance, the GLMNET implementation may not converge [62,70]. A case in point is that [71] provides two numerical experiments in which GLMNET does not converge.

Other coordinate descent algorithms have been developed to compute solutions to non-smooth MaxEnt models directly (see, e.g., [64,72]) but they are also generally non-parallelizable and often lack robustness and good convergence properties. Finally, another issue is that many coordinate descent algorithms, including GLMNET, were designed for sparse MaxEnt models (e.g., the elastic net and group lasso regularized MaxEnt models described in Section 2.3) and those algorithms depend on the sparsity of the model to converge quickly [54]. It would be desirable to have fast optimization methods for when the feature mapping (13) is dense, as this often occurs in practice.

First-order methods. First-order optimization algorithms such as the forward–backward splitting algorithm are popular because they are robust and can provably compute ϵ-approximate solutions of (3) with an optimal rate of convergence. However, as discussed in detail at the end of the previous subsection, achieving this rate of convergence requires fine-tuning the stepsize parameters using an accurate estimate of the largest singular value of the feature mapping (13), and this estimate is typically computationally expensive for large matrices. This problem makes classical first-order optimization algorithms inefficient and impractical in estimating probability densities from large-scale MaxEnt models.

Summary. The current state-of-the-art algorithms for estimating probability densities from MaxEnt models either produce unreliable numerical results, lack scalable parallelism, or scale poorly in size. These shortcomings in terms of robustness and efficiency make it challenging to use non-smooth, large-scale MaxEnt models in applications without access to adequate and costly computational resources. The next section presents novel, efficient, and robust accelerated NPDHG optimization methods that address these shortcomings.

4. Main Results

We describe here our approach for computing solutions to the generalized MaxEnt problem (3) using accelerated NPDHG optimization methods [38]. In addition to the standing assumptions on the potential function H^*, we assume we can compute efficiently the proximal operator of the convex conjugate of the potential function

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^m}\left\{\lambda H(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2\right\},$$

for every $\lambda > 0$ and $\hat{\boldsymbol{w}} \in \mathbb{R}^m$. These assumptions are satisfied for most potential functions used in practice, including for the potential functions described in Section 2.3.

4.1. Methodology

We start with the saddle-point formulation (4) of the generalized MaxEnt estimation problem (3). Based on the work two of the authors provided in [38], we propose to split the infimum and supremum in the saddle-point problem (4) using an iterative NPDHG scheme that alternates between a nonlinear proximal ascent step based on the Kullback–Leibler divergence and a linear proximal descent step.

More precisely, let $\tau_0 > 0$ and $\sigma_0 > 0$ be two stepsize parameters satisfying the inequality

$$\tau_0\sigma_0\left(\max_{j\in\{1,\dots,n\}}\|\boldsymbol{\Phi}(j)\|_2^2\right) \leqslant 1, \tag{14}$$

let $\theta_0 = 0$, let $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 \in \text{int dom } H$, let $\boldsymbol{z}_0 \in \mathbb{R}^m$, and define the initial probability distribution $p_0 \in \Delta_n$ through the constant $\boldsymbol{z}_0$ via

$$p_0(j) = \frac{p_{\text{prior}}(j)e^{\langle \boldsymbol{z}_0, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \boldsymbol{z}_0, \boldsymbol{\Phi}(j)\rangle}} \tag{15}$$

for each $j \in \{1,\dots,n\}$. Our proposed NPDHG iterative scheme computes the iterates

$$
\begin{aligned}
p_{k+1} &= \arg\min_{p\in\Delta_n}\left\{\tau_k D_{\text{KL}}(p \parallel p_{\text{prior}}) - \tau_k\langle \boldsymbol{w}_k + \theta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}), \mathbb{E}_p[\boldsymbol{\Phi}]\rangle + D_{\text{KL}}(p \parallel p_k)\right\} \\
\hat{\boldsymbol{w}}_{k+1} &= \boldsymbol{w}_k + \sigma_k(\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p_{k+1}}[\boldsymbol{\Phi}]), \\
\boldsymbol{w}_{k+1} &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^m}\left\{t\sigma_k H(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{k+1}\|_2^2\right\}, \\
\theta_{k+1} &= 1/\sqrt{1 + \tau_k}, \quad \tau_{k+1} = \theta_{k+1}\tau_k \quad \text{and} \quad \sigma_{k+1} = \sigma_k/\theta_{k+1}.
\end{aligned} \tag{16}
$$

According to ([38], Proposition 4.1), the sequence of iterates p_k converges strongly to the unique solution of the generalized MaxEnt problem (3). Moreover, for any $t \geqslant 0$ and a given tolerance level $\epsilon > 0$, the scheme (16) provably computes an ϵ-approximate solution $p^\epsilon(t)$ of the generalized MaxEnt model (3) in $O(1/\sqrt{\epsilon})$ time. This rate of convergence is, without further smoothness assumptions on the potential function H^*, the best achievable rate of convergence with respect to the Nesterov class of optimal first-order methods [39].

The key element in this scheme is the choice of the Kullback–Leibler divergence as a nonlinear proximal step in the first line of (16). We use it for two reasons: First, because the Kullback–Leibler divergence already appears in the saddle-point problem (4). This allows us to compute p_{k+1} explicitly. Indeed, thanks to the choice of initial probability distribution (15), we have

$$p_{k+1}(j) = \frac{p_{\text{prior}}(j)e^{\langle \boldsymbol{z}_{k+1}, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\text{prior}}(j)e^{\langle \boldsymbol{z}_{k+1}, \boldsymbol{\Phi}(j)\rangle}} \text{ with } \boldsymbol{z}_{k+1} = \frac{1}{1 + \tau_k}(\boldsymbol{z}_k + \tau_k(\boldsymbol{w}_k + \theta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}))) \tag{17}$$

for each $j \in \{1,\dots,n\}$. See Appendix A for the derivation of (17).

Second, because the Kullback–Leibler divergence is 1-strongly convex with respect to the ℓ_1 norm, that is,

$$D_{\text{KL}}(p \parallel p_k) \geqslant \frac{1}{2}\|p - p_k\|_1^2.$$

This fact follows from a fundamental result in information theory known as Pinsker's inequality [47–51]. In particular, this means that scheme (16) alternates between solving a strongly convex problem over the space $(\mathbb{R}^n, \|\cdot\|_1)$ and a concave problem over the space

$(\mathbb{R}^m, \|\cdot\|_2)$. The choice of these spaces is significant, for the induced operator norm $\|\cdot\|_{\mathrm{op}}$ of the linear operator A defined in (13) becomes

$$\|A\|_{\mathrm{op}} = \sup_{\|p\|_1=1} \|Ap\|_2 = \max_{j\in\{1,\dots,n\}} \|\boldsymbol{\Phi}(j)\|_2.$$

This operator norm offers two crucial advantages: First, it can be computed in optimal $\Theta(mn)$ time, or better if the features $\{\boldsymbol{\Phi}(j)\}_{j=1}^n$ have structure, e.g., if they are sparse. This means that the stepsize parameters τ_0 and σ_0 of the NPDHG scheme can be computed in (14), with equality, in optimal $\Theta(mn)$ time. This is important because, in classical first-order optimization methods, we typically require a precise estimate of the largest singular value of the feature mapping (13), namely the number

$$\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|Ap\|_2,$$

to fine-tune the stepsize parameters to gain an optimal convergence rate. However, as discussed in detail in Section 3.1, this estimate is computationally expensive for large matrices due to its prohibitive cubic cost of $O(\min(m^2n, mn^2))$ operations. In contrast, our NPDHG scheme (16) does not suffer from this computational bottleneck and therefore scales well. Second, the operator norm $\|A\|_{\mathrm{op}}$ can be significantly smaller than the estimate $\|A\|_2$, hence allowing for bigger stepsize parameters to further speed up convergence. (An easy calculation yields $\max_{\|\boldsymbol{p}\|_2=1}\|A\boldsymbol{p}\|_2 \geqslant \max_{j\in\{1,\dots,n\}}\|\boldsymbol{\Phi}(j)\|_2$.)

Summary. To solve the generalized MaxEnt estimation problem (3) and its dual problem (5), let $\tau_0 > 0$ and $\sigma_0 > 0$ be two stepsize parameters satisfying inequality (14), let $\theta_0 = 0$, let $w_{-1} = w_0 \in \mathrm{int\ dom}\ H$, let $z_0 \in \mathbb{R}^m$, and define an initial probability distribution p_0 through z_0 via (15). Then, compute the iterates

$$
\begin{aligned}
z_{k+1} &= (z_k + \tau_k(w_k + \theta_k(w_k - w_{k-1})))/(1 + \tau_k), \\
p_{k+1}(j) &= \frac{p_{\mathrm{prior}}(j)e^{\langle z_{k+1}, \boldsymbol{\Phi}(j)\rangle}}{\sum_{j=1}^n p_{\mathrm{prior}}(j)e^{\langle z_{k+1}, \boldsymbol{\Phi}(j)\rangle}} \quad \text{for } j \in \{1,\dots,n\}, \\
\hat{w}_{k+1} &= w_k + \sigma_k(\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p_{k+1}}[\boldsymbol{\Phi}]), \\
w_{k+1} &= \operatorname*{arg\,min}_{w\in\mathbb{R}^m}\left\{ t\sigma_k H(w) + \frac{1}{2}\|w - \hat{w}_{k+1}\|_2^2 \right\}, \\
\theta_{k+1} &= 1/\sqrt{1 + \tau_k}, \quad \tau_{k+1} = \theta_{k+1}\tau_k \quad \text{and} \quad \sigma_{k+1} = \sigma_k/\theta_{k+1},
\end{aligned}
\tag{18}
$$

until convergence is achieved. As all parameters and updates can be computed in $O(mn)$ time, for any $t \geqslant 0$ and a given tolerance level $\epsilon > 0$, the overall complexity for computing an ϵ-approximate solution $p^\epsilon(t)$ is $O(mn/\sqrt{\epsilon})$.

4.2. Algorithm for Smooth Potential Functions

The iterative scheme (18) does not require the potential function H^* to be smooth. If, however, the potential function H^* is γ_{H^*}-smooth (equivalently, H is $\frac{1}{\gamma_{H^*}}$-strongly convex) for some $\gamma_{H^*} > 0$, then we can modify the NPDHG iterative scheme (18) to achieve a linear rate of convergence. More precisely, let $t > 0$ and introduce the stepsize parameters

$$\theta = 1 - \frac{t}{2\gamma_{H^*}\|A\|_{\mathrm{op}}^2}\left(\sqrt{1 + \frac{4\gamma_{H^*}\|A\|_{\mathrm{op}}^2}{t}} - 1\right), \quad \tau = \frac{1-\theta}{\theta} \quad \text{and} \quad \sigma = \frac{\gamma_{H^*}\tau}{t}. \tag{19}$$

Let $w_{-1} = w_0 \in \text{int dom } H$, let $z_0 \in \mathbb{R}^m$, and define p_0 through z_0 via (15). Then, the explicit NPDHG iterative scheme is

$$
\begin{aligned}
z_{k+1} &= (z_k + \tau(w_k + \theta(w_k - w_{k-1})))/(1 + \tau), \\[4pt]
p_{k+1}(j) &= \frac{p_{\text{prior}}(j)e^{\langle z_{k+1}, \Phi(j) \rangle}}{\sum_{j=1}^{n} p_{\text{prior}}(j)e^{\langle z_{k+1}, \Phi(j) \rangle}} \quad \text{for } j \in \{1, \ldots, n\}, \\[4pt]
\hat{w}_{k+1} &= w_k + \sigma(\mathbb{E}_{\hat{\mathcal{D}}}[\Phi] - \mathbb{E}_{p_{k+1}}[\Phi]), \\[4pt]
w_{k+1} &= \operatorname*{arg\,min}_{w \in \mathbb{R}^m}\left\{ t\sigma H(w) + \frac{1}{2}\|w - \hat{w}_{k+1}\|_2^2 \right\}.
\end{aligned}
\tag{20}
$$

According to ([38], Proposition 4.3), the sequences of iterates p_k and w_k converge strongly to the unique solution of the generalized MaxEnt estimation problem (3) and its dual problem (5). Moreover, for any $t > 0$ and a given tolerance level $\epsilon > 0$, this scheme provably computes an ϵ-approximate solution $p^\epsilon(t)$ of the generalized MaxEnt estimation problem (3) in $O(\log(1/\epsilon))$. This rate of convergence is the best achievable rate of convergence with respect to the Nesterov class of optimal first-order methods [39]. As all parameters and updates can be computed in $O(mn)$ time, the overall complexity for computing an ϵ-approximate solution $p^\epsilon(t)$ is $O(mn \log(1/\epsilon))$.

5. Application to Wildfire Science

To illustrate the efficiency of our novel algorithms on large-scale problems, we present here an application to wildfire science. The problem at hand is to combine fire occurrence data with ecological data in a fixed geographical region to estimate the probability of fire occurrences as a function of ecological features. MaxEnt models achieve this goal by translating fire occurrence and ecological data into probabilities of fire occurrences and ecological features. This approach closely mirrors how MaxEnt models are used for modeling species' geographic distributions [19,20,22,24–28]. Another related goal is to identify what ecological features correlate most with fire occurrences. This can be achieved using a sparse MaxEnt model, e.g., an elastic net or group lasso regularized MaxEnt model, to identify ecological features correlating significantly with fire occurrences.

For this application, we use the Western US MTBS-Interagency (WUMI) wildfire data set [45], which we describe in Section 5.1 below. We formulate the problem of combining the fire occurrence and ecological data from the WUMI wildfire data set into a MaxEnt estimation problem in Section 5.2. Using this data, we then fit the elastic net, (non-overlapping) group lasso, and ℓ_∞ MaxEnt models for a large number of hyperparameters weighting the regularization. We detail this fitting procedure and the explicit NPDHG algorithms for these MaxEnt models in Section 5.3. Following this, we compare in Section 5.4 the running times required to fit the aforementioned MaxEnt models using our NPDHG algorithms with the forward–backward splitting algorithm [40,65] and the STRUCTMAXENT2 coordinate descent algorithm of [41]. Finally, in Section 5.5, we interpret the results obtained from fitting the aforementioned MaxEnt models to the WUMI wildfire data set.

5.1. WUMI Wildfire Data Set

The Western US MTBS-Interagency wildfire data set [45] consists of all fires ($\geqslant 1$ km^2) from multiple state and federal agencies, supplemented by satellite observations of large fires ($\geqslant 4$ km^2) from the Monitoring Trends in Burn Severity (MTBS) program, in the continental United States west of $103°$ W longitude. For this application, we extracted all wildfires from the WUMI data set that occurred between 1984 and 2020 inclusive (accessed 18 May 2023). The locations of all fires used are shown in Figure 1.

Next, following the procedure outlined in [44], we overlay the fire locations on a 12 km $\times$ 12 km grid to construct a data frame of prevailing climate, vegetation, topographic, and human-related features for each fire. Altogether we include a total of 35 potential

fire-related features from various sources; we provide a summary of all features used in our analysis along with their sources in Appendix B. For each grid cell, we also provide an index identifying its Environmental Protection Agency (EPA) Level III ecoregion. Defined on the basis of roughly homogeneous climate and vegetation patterns, ecoregions are commonly used in the wildfire science literature to identify climate–fire relationships at a coarse spatial scale [29,73]. The full WUMI wildfire data set with associated fire-related features for all months from 1984 to 2020 is publicly available as part of the Stochastic Machine Learning for wildfire activity in the western US (SMLFire1.0) package [44].
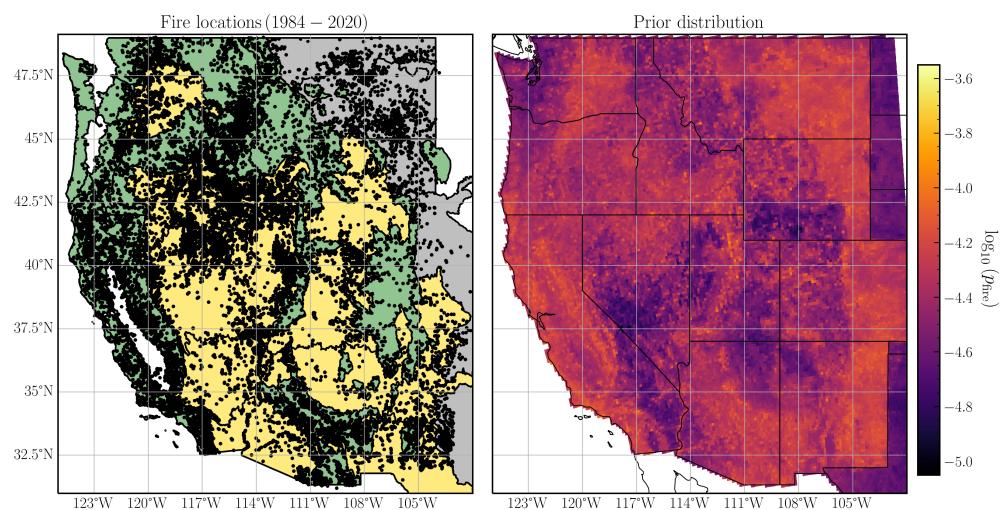

Figure 1. Wildfire activity in the western United States from 1984 to 2020. (**Left**) Fire locations of all fires (black dots) in the Western US MTBS-Interagency (WUMI) data set; also shown are three ecological divisions characterized by their primary vegetation type—forests (green), deserts (yellow), and plains (gray). (**Right**) Prior distribution indicating mean fire probability across all calendar months.

5.2. Data Preprocessing

We construct a weakly informative prior [74,75] for incorporating existing knowledge of wildfire conducive environmental conditions in our MaxEnt model. First, we prepare a training data set of 10,000 fire occurrences and absences chosen randomly across all months between 1984 and 2020 correlated with the values of 35 fire-related features described in the previous section. The features are averaged over each calendar month (i.e., January, February, ...) between 1984 and 2020. We then apply min–max scaling to each feature, ensuring that all features lie in the same range. Second, we construct two Random Forest (RF) models, one each for fires in dry (May–September) and wet (October–April) seasons respectively. We withhold 20% of the training data to tune the model hyperparameters, such that the optimal hyperparameters ensure a trade-off between model precision and recall. Next, we predict the fire probability for all grid cells with fire-related features using either the trained wet or dry season RF model depending on the calendar month. Since the MaxEnt algorithm assumes a presence-only framework, that is, the absence of fires in a grid cell does not imply a non-zero fire probability, we impute grid cells without a fire with probability $p_{\text{nfire}} = \min(p_{\text{fire}})/10$. Last, we normalize the prior probability distribution to ensure that the fire probability across all grid cells sums up to one. For convenience, we represent the prior distribution as the normalized mean of 12 monthly fire probabilities predicted by the RF models in Figure 1.

To construct the empirical distribution $\hat{\mathcal{D}}$, we divide the study region into its 18 different EPA level III ecoregions and weigh the relative frequencies of fire among the ecoregions using the following strategy. For each ecoregion $r \in \{1, \ldots, 18\}$, let $n_{\text{r,total}}$ denote the total number of grid cells in ecoregion r and let $n_{\text{r,fire}}$ denote the total number grid cell in ecoregion r where at least one fire was recorded. In addition, let $Z = \sum_{r=1}^{18} \frac{n_{\text{r,fire}}}{n_{\text{r,total}}}$ denote the sum

of the proportions of grid cells where at least one fire was recorded among all ecoregions. Then, we compute the empirical probability at cell $j \in \{1, \ldots, n\}$, $\hat{\mathcal{D}}(j)$, as follows:

$$\hat{\mathcal{D}}(j) = \frac{1}{Z} \begin{cases} 0 & \text{if no fire was recorded in grid cell } j, \\ 1/n_{\text{r,total}} & \text{if at least one fire was recorded in grid } j \text{ and the grid} \\ & \text{cell } j \text{ belongs to ecoregion } r \end{cases} \tag{21}$$

This construction gives the empirical distribution more weight to ecoregions where fires are more frequent and widespread.

5.3. Fitting Procedure and Algorithmic Setup

For the analysis, we fit six different MaxEnt models to the wildfire data: four elastic net MaxEnt models with parameters $\alpha = \{0.95, 0.4, 0.15, 0.05\}$, a non-overlapping group lasso MaxEnt model, and an ℓ_∞-regularized MaxEnt model. For each model, we fit a regularization path of 141 hyperparameters as follows:

$$t^{(l)} = \begin{cases} (1 - l/100)t^{(0)} & \text{from } l = 0 \text{ to } 50 \\ (0.5 - (l - 50)/200)t^{(0)} & \text{from } l = 51 \text{ to } 140, \end{cases}$$

where $t^{(0)}$ depends on the MaxEnt model and corresponds to the smallest hyperparameter for which the primal and dual solutions are equal to the prior distribution and zero. More precisely, for each model we set $(p^{(0)}, \boldsymbol{w}^{(0)}) = (p_{\text{prior}}, \boldsymbol{0})$ and compute the sequence of solutions $\{(p^{(l)}, \boldsymbol{w}^{(l)})\}_{l=1}^{140}$ of the corresponding MaxEnt primal and dual problems (7) and (8) with hyperparameter $t = t^{(l)}$.

We chose these MaxEnt models to study the impact of the regularization on the primal and dual solutions as a function of the hyperparameters. In particular, for the elastic net and non-overlapping group lasso MaxEnt models, we are interested in identifying the set of features that are selected or discarded as a function of the sequence of hyperparameters $\{t^{(l)}\}_{l=0}^{140}$. In the parts below, we describe the value of $t^{(0)}$ and the NPDHG algorithm used for each model.

5.3.1. Elastic Net MaxEnt Models

For these MaxEnt models, the smallest hyperparameter for which the solutions to the primal and dual problems (7) and (8) are equal to the prior distribution and zero is

$$t^{(0)} = \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p_{\text{prior}}}[\boldsymbol{\Phi}] \right\|_\infty / \alpha.$$

This follows from the optimality condition

$$\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^s}[\boldsymbol{\Phi}] - t(1 - \alpha)\boldsymbol{w}^s \in t\alpha \partial \|\cdot\|_1 \boldsymbol{w}^s \tag{22}$$

and computing the smallest parameter t for which (22) is satisfied at $(p^s, \boldsymbol{w}^s) = (p_{\text{prior}}, \boldsymbol{0})$.

For the elastic net MaxEnt model with $\alpha = 0.95$, we use the NPDHG algorithm (18) (with sublinear convergence rate), while, for the models with $\alpha = \{0.4, 0.15, 0.05\}$, we use the NPDHG algorithm (20) (with linear convergence rate). Starting from $l = 1$, we compute the pair of solutions $(p^{(l)}, \boldsymbol{w}^{(l)})$ to the primal and dual problems (7) and (8) at hyperparameter $t = t^{(l)}$ using the previously computed pair of solutions $(p^{(l-1)}, \boldsymbol{w}^{(l-1)})$ by setting the initial vectors $\boldsymbol{z}_0 = \boldsymbol{w}_0 = \boldsymbol{w}^{(l-1)}$ in both NPDHG algorithms. For the stepsize parameters, we set $\theta_0 = 0$, $\tau_0 = 2$, and $\sigma_0 = 1/2\|\|A\|\|_{\text{op}}^2$ in (18), and we set θ, τ, and σ according to the formulas in (19) with $\gamma_{H^*} = 1 - \alpha$ in (20). We compute the update

w_{k+1} in both NPDHG algorithms using the classical soft thresholding operator [66,76,77]. Specifically, for any $\lambda > 0$, $\hat{w} \in \mathbb{R}^m$, and $i \in \{1, \ldots, m\}$ we have

$$\left[\underset{w \in \mathbb{R}^m}{\arg\min} \left\{ \lambda \|w\|_1 + \frac{1}{2} \|w - \hat{w}\|_2^2 \right\} \right]_i \equiv [\mathrm{shrink}_1(\hat{w}, \lambda)]_i = \begin{cases} [\hat{w}]_i - \lambda & \text{if } [\hat{w}]_i > \lambda, \\ 0 & \text{if } |[\hat{w}]_i| \leqslant \lambda, \\ [\hat{w}]_i + \lambda & \text{if } [\hat{w}]_i < -\lambda, \end{cases}$$

and so, for every $\lambda > 0$, $\alpha \in [0, 1]$, $\hat{w} \in \mathbb{R}^m$ and $i \in \{1, \ldots, m\}$ we have

$$\left[\underset{w \in \mathbb{R}^m}{\arg\min} \left\{ \lambda \left(\alpha \|w\|_1 + \frac{(1-\alpha)}{2} \|w\|_2^2 \right) + \frac{1}{2} \|w - \hat{w}\|_2^2 \right\} \right]_i = \frac{[\mathrm{shrink}_1(\hat{w}, \lambda\alpha)]_i}{1 + \lambda(1-\alpha)}.$$

Finally, we let the NPDHG algorithms run for at least 40 iterations before checking for convergence. We stop the NPDHG algorithms when the optimality condition (22) is satisfied within some tolerance 10^{-5}:

$$\left\| \mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] - \mathbb{E}_{p_k}[\mathbf{\Phi}] - t(1-\alpha)w_k \right\|_\infty \leqslant t\alpha(1 + 10^{-5}).$$

5.3.2. Non-Overlapping Group Lasso MaxEnt Model

For this MaxEnt model, we divide the features into five disjoint groups of features, as described in Appendix B. Then, the smallest hyperparameter for which the solutions to the primal and dual problems (7) and (8) are equal to the prior distribution and zero is

$$t^{(0)} = \max_{g \in \{1, \ldots, 5\}} \left\{ \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}_g] - \mathbb{E}_p[\mathbf{\Phi}_g] \right\|_{2,g} / \sqrt{m_g} \right\},$$

where m_g is the number of features in the gth group. This follows from the optimality condition

$$\mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] - \mathbb{E}_{p^s}[\mathbf{\Phi}] \in \bigcup_{g=1}^{5} \left\{ t\sqrt{m_g} \partial \|\cdot\|_{2,g} \right\} w_g^s \tag{23}$$

and computing the smallest parameter t for which (23) is satisfied at $(p^s, w^s) = (p_{\mathrm{prior}}, \mathbf{0})$.

For this model, we use the NPDHG algorithm (18). Starting from $l = 1$, we compute the pair of solutions $(p^{(l)}, w^{(l)})$ to the primal and dual problems (10) and (9) at hyperparameter $t = t^{(l)}$ using the previously computed pair of solutions $(p^{(l-1)}, w^{(l-1)})$ by setting the initial vectors $z_0 = w_0 = w^{(l-1)}$. For the stepsize parameters, we set $\theta_0 = 0$, $\tau_0 = 2$, and $\sigma_0 = 1/2\|\|A\|\|_{\mathrm{op}}^2$. We compute the update w_{k+1} in (18) using the following proximal operator formula: for every group $g \in \{1, \ldots, 5\}$, $\lambda > 0$ and $\hat{w} \in \mathbb{R}^m$,

$$\underset{w \in \mathbb{R}^m}{\arg\min} \left\{ \lambda \sqrt{m_g} \|w\|_{2,g} + \frac{1}{2} \|w - \hat{w}\|_{2,g}^2 \right\} = \max\left(0, 1 - \lambda \sqrt{m_g} / \|\hat{w}\|_2 \right) \hat{w}.$$

Finally, we let the NPDHG algorithms run for at least 40 iterations before checking for convergence. We stop the NPDHG algorithms when the optimality condition (23) is satisfied within some tolerance 10^{-5}:

$$\max_{g \in \{1, \ldots, 5\}} \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] - \mathbb{E}_{p_k}[\mathbf{\Phi}] \right\|_{2,g} \leqslant t(1 + 10^{-5}).$$

5.3.3. ℓ_∞-Regularized MaxEnt Model

For this MaxEnt model, the smallest hyperparameter for which the solutions to the primal and dual problems (7) and (8) are equal to the prior distribution and zero is

$$t^{(0)} = \left\| \mathbb{E}_{\hat{\mathcal{D}}}[\mathbf{\Phi}] - \mathbb{E}_{p_{\mathrm{prior}}}[\mathbf{\Phi}] \right\|_1.$$

This follows from the optimality condition

$$\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p^s}[\boldsymbol{\Phi}] \in t\partial\|\cdot\|_{\infty}(\boldsymbol{w}^s) \tag{24}$$

and computing the smallest parameter t for which (24) is satisfied at $(p^s, \boldsymbol{w}^s) = (p_{\text{prior}}, \boldsymbol{0})$.

For this model, we use the NPDHG algorithm (18). Starting from $l = 1$, we compute the pair of solutions $(p^{(l)}, \boldsymbol{w}^{(l)})$ to the primal and dual problems (11) and (12) at hyperparameter $t = t^{(l)}$ using the previously computed pair of solutions $(p^{(l-1)}, \boldsymbol{w}^{(l-1)})$ by setting the initial vectors $\boldsymbol{z}_0 = \boldsymbol{w}_0 = \boldsymbol{w}^{(l-1)}$. For the stepsize parameters, we set $\theta_0 = 0$, $\tau_0 = 2$, and $\sigma_0 = 1/2\||\boldsymbol{A}|\|_{\text{op}}^2$. Finally, we compute the update $\boldsymbol{w}_{k+1}$ in (18) using Moreau's decomposition ([78], Theorem 3.2.5): for every $\lambda > 0$ and $\hat{\boldsymbol{w}} \in \mathbb{R}^m$,

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^m}\left\{\lambda\|\boldsymbol{w}\|_{\infty} + \frac{1}{2}\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2\right\} = \hat{\boldsymbol{w}} - \arg\min_{\|\boldsymbol{w}\|_1 \leqslant \lambda}\frac{1}{2}\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2.$$

The second term on the right amounts to projecting $\hat{\boldsymbol{w}}$ on the ℓ_1 ball of radius λ. There are fast algorithms for doing do; we use Algorithm 1 described in [79]. Finally, we let the NPDHG algorithms run for at least 40 iterations before checking for convergence. We stop the NPDHG algorithms when the optimality condition (24) is satisfied within some tolerance 10^{-5}:

$$\left\|\mathbb{E}_{\hat{\mathcal{D}}}[\boldsymbol{\Phi}] - \mathbb{E}_{p_k}[\boldsymbol{\Phi}]\right\|_1 \leqslant t(1 + 10^{-5}).$$

5.4. Comparison of Timings

In this section, we compare the run times of our NPDHG algorithms with two state-of-the-art optimization algorithms for solving non-smooth MaxEnt models: the forward–backward splitting algorithm (specifically, Algorithm 5 in [65]; see also [40,66]) and the STRUCTMAXENT2 coordinate descent algorithm from [41,64]. All numerical experiments were performed on a single core Intel(R) Core(TM) i7-10750H CPU @ 2.60 GHz.

We initially chose the GLMNET implementation [26,33] in MATLAB 2023b over STRUCTMAXENT2 for the numerical comparisons, but we found that GLMNET produced unreliable numerical solutions when compared to both the NPDHG and forward–backward splitting algorithms. We also tried using GLMNET's implementation in the R language, but to no avail. We think this problem arises because, as discussed in Section 3.2, the GLMNET algorithm approximates a MaxEnt model as a logistic regression model and then invokes a coordinate descent method tailored to logistic regression to approximate the solution to the MaxEnt model. Our observations suggest this approach does not work well for our data set. In contrast, we found that the STRUCTMAXENT2 algorithm produced correct numerical results.

For the forward–backward splitting algorithm, the stepsize parameters were set to $1/\|\boldsymbol{A}\|_2$ and 1 (corresponding to τ and t_0 in Algorithm 5 of [65]) and, for computing the pair of solutions $(p^{(l)}, \boldsymbol{w}^{(l)})$ at hyperparameter $t^{(l)}$, the initial iterate was set to $\boldsymbol{w}^{(l-1)}$. In addition, for the elastic net MaxEnt models, the acceleration quantity $q = (1 - \alpha)t^{(l)}/(\|\boldsymbol{A}\|_2 + (1 - \alpha)t^{(l)})$ was employed. We used the stopping criteria of the NPDHG algorithms for the forward–backward splitting algorithms. For the coordinate descent algorithm, we modified the STRUCTMAXENT2 algorithm from ([41], pp. 305–306) to make it applicable to the elastic net penalty. For computing the pair of solutions $(p^{(l)}, \boldsymbol{w}^{(l)})$ at hyperparameter $t^{(l)}$, the initial iterate was set to $\boldsymbol{w}^{(l-1)}$. We did not use the STRUCTMAXENT2 algorithm for the non-overlapping group lasso or ℓ_{∞}-regularized MaxEnt models, as it was not designed for these MaxEnt models. We used the stopping criteria of the NPDHG algorithms for the STRUCTMAXENT2 algorithm.

Table 1 shows the average timings for computing the entire regularization path of the WUMI wildfire data set using the coordinate descent, forward–backward splitting, and NPDHG algorithms. All timings were averaged over five runs and, for the forward–

backward splitting and NPDHG algorithms, they include the time required to compute all the stepsize parameters. All algorithms were implemented in MATLAB.

Table 1. Timings results (in seconds) for fitting the MaxEnt models described in Section 5.3. All timings are averaged over five runs.

	STRUCTMAXENT2	Forward–Backward Splitting	NPDHG
Elastic net ($\alpha = 0.95$)	5562.19	4208.01	365.55
Elastic net ($\alpha = 0.40$)	1018.73	1407.22	113.53
Non-overlapping group lasso	N/A	3036.38	278.14
ℓ_∞-regularization	N/A	2534.65	289.98

The NPDHG algorithm outperformed both the forward–backward splitting and coordinate descent algorithms by at least one order of magnitude. In particular, we observed that the NPDHG algorithm required far fewer iterations to achieve convergence compared to both the forward–backward splitting and STRUCTMAXENT2 algorithms. This difference is because the stepsize parameters for the NPDHG algorithm were much larger compared to either the forward–backward splitting or the STRUCTMAXENT2 algorithm. Indeed, the stepsize parameters for the NPDHG and forward–backward splitting algorithms are inversely proportional to the norms $\|A\|_{\mathrm{op}}$ and $\|A\|_2$, and for the wildfire data set these were $\|A\|_{\mathrm{op}} \approx 3.30$ and $\|A\|_2 \approx 854.08$. Thus, larger stepsize parameters were permitted thanks to the Kullback–Leibler divergence term in the NPDHG algorithm, enabling a major speedup gain.

5.5. Analysis of the MaxEnt Regularization Paths and Estimated Fire Probabilities

As a final validation step for the NPDHG algorithm, we use the fitted MaxEnt models to compute the normalized mean fire probability in each grid cell for all calendar months between 1984 and 2020. The probabilities are visualized in Figures 2 and 3. In each case, we have chosen the prior distribution as a benchmark for our spatial plots of fire probability.

The spatial fire probabilities for $\alpha = \{0.95, 0.40, 0.15, 0.05\}$ are shown in Figure 2. The range of α values roughly corresponds to varying the regularization from a purely l_1 norm ($\alpha = 1$) to a purely l_2 norm ($\alpha = 0$). For each value of α, we also consider the evolution of the spatial fire probability as we vary the hyperparameter t along the regularization path, or equivalently include additional features while fitting the MaxEnt model to wildfire data. Broadly, we observe that, for a fixed value of $t/t_{\max}$, the ratio of predicted to prior fire probability decreases in sharpness as α decreases. On the other hand, for a fixed α, decreasing $t/t_{\max}$ values enables the model to make sharper distinctions between grid cells with high and low fire probability as evidenced by the sharper contrast between the prior and predicted fire probabilities. We also note a similar pattern in Figure 3 for the MaxEnt models with non-overlapping group lasso (corresponding roughly to the elastic net case with $\alpha = 1$) and the l_∞-MaxEnt models, all of which converge to the empirical distribution quicker than any of the elastic net cases.

In Figure 4, we show the cumulative number of non-zero coefficients at fixed intervals along the regularization path for different α values. The plot helps in visualizing the $t/t_{\max}$ values at which new features are introduced in the elastic net MaxEnt model, with the dashed vertical lines indicating the first time a feature from a new group is selected. Across all α values, we find that features appear in the same order, with fire weather features being selected first, followed by topography, vegetation, human, and antecedent features. We tabulate the final set of non-zero features at the end of the regularization path for $\alpha = \{0.95, 0.05\}$ across various groups in Table 2. These selected features are in good agreement with physical models of wildfire occurrence [42,43] as well as previous statistical analyses of wildfire drivers [29–31,44].
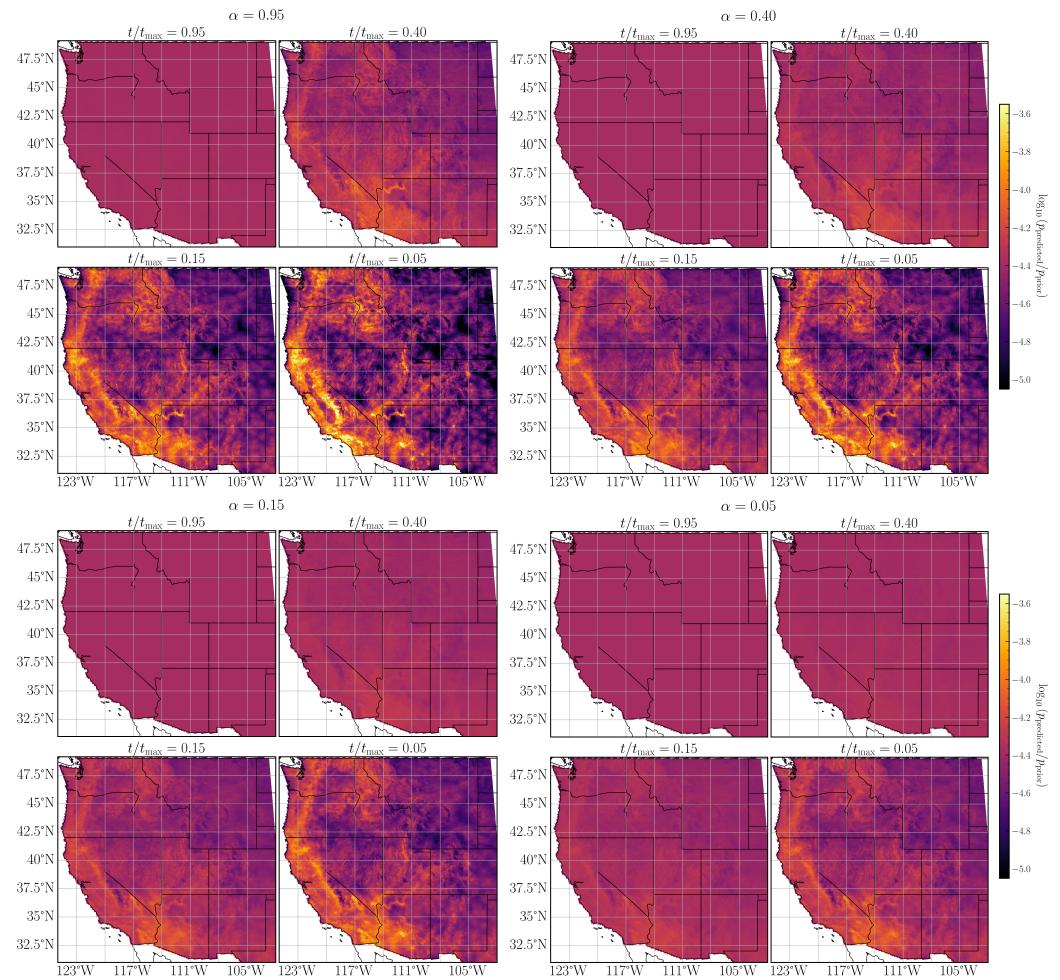
Figure 2. Spatial probability plot for different hyperparameter values with elastic net penalty parameter $\alpha = \{0.95, 0.40, 0.15, 0.05\}$.
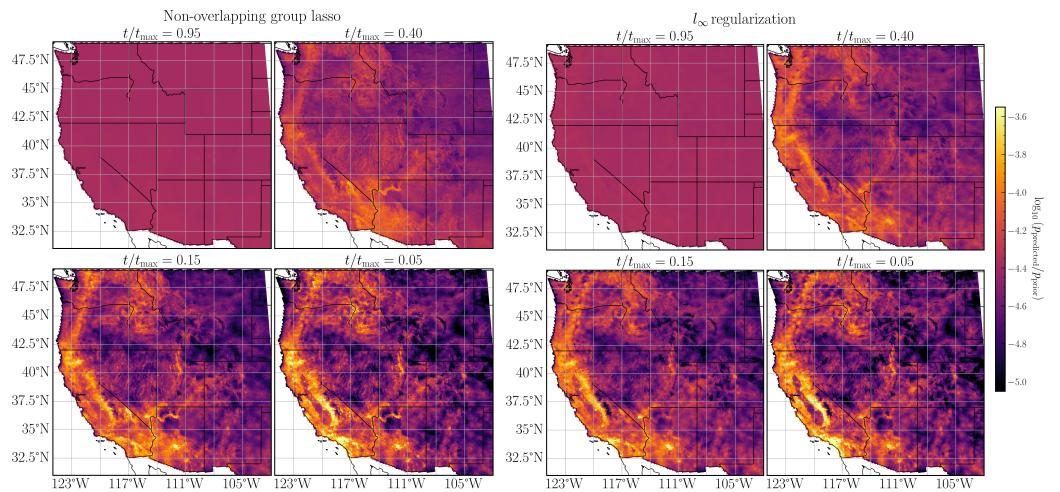

Figure 3. Same as Figure 2 but for (**left**) the non-overlapping group lasso with $\alpha = 1$ and (**right**) the l_∞ MaxEnt models, respectively.

Table 2. List of non-zero features at the end of the regularization path for two elastic net penalty parameters, organized by different feature groups. See Appendix B for additional description of the selected features.

Feature Group	$\alpha = 0.95$	$\alpha = 0.05$	
Fire weather	Tmax	Tmax	VPD
	Prec	Prec	Tmin
	Wind	Wind	VPD_{max3}
	FM1000	FM1000	VPD_{max7}
	$\text{Tmin}_{\text{max3}}$	$\text{Tmin}_{\text{max3}}$	$\text{Tmax}_{\text{max3}}$
	Lightning	Lightning	$\text{Tmax}_{\text{max7}}$
		$\text{Tmin}_{\text{max7}}$	SWE_{max}
Topography	Slope	Slope	Southness
	Southness		
Vegetation	Grassland	Biomass	Forest
		Shrub	
Human	Camp_dist	Camp_dist	Road_dist
	Camp_num	Camp_num	
	Urban	Urban	
	Popdensity	Popdensity	
Antecedent	$\text{AvgPrec}_{\text{2mo}}$	$\text{AvgPrec}_{\text{2mo}}$	$\text{AntPrec}_{\text{lag2}}$
	$\text{AvgVPD}_{\text{3mo}}$	$\text{AntPrec}_{\text{lag1}}$	
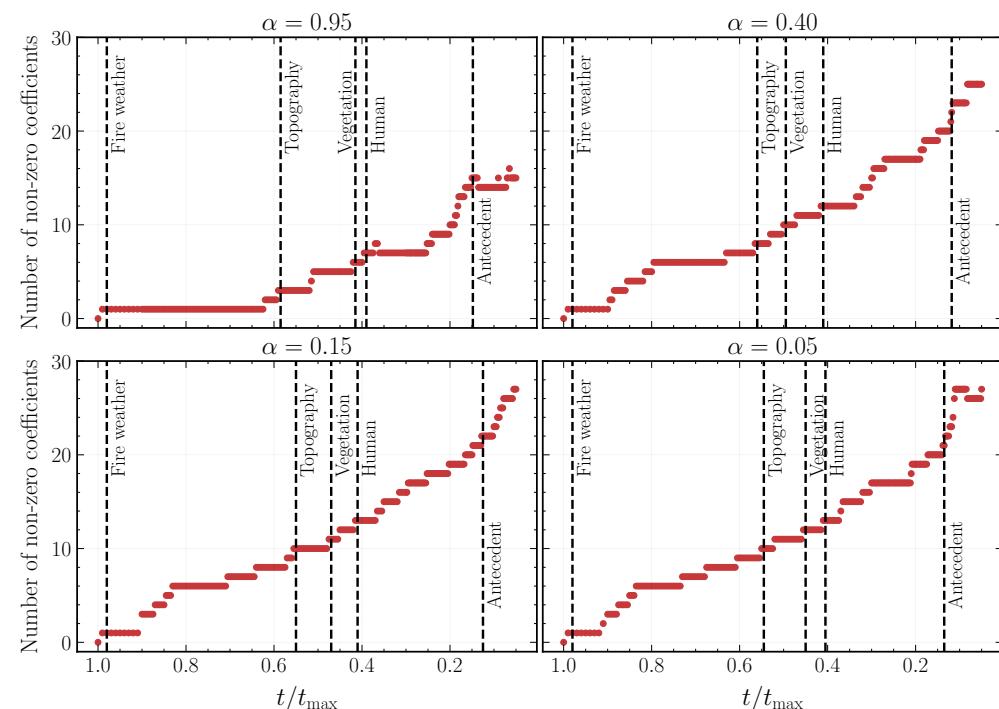
Figure 4. Number of non-zero coefficients along the regularization path plots for elastic net penalty parameter $\alpha = \{0.95, 0.40, 0.15, 0.05\}$. The dashed vertical lines highlight the $t/t_{\max}$ value at which the first feature of the group indicated by inset text is selected.

6. Discussion

In this paper, we have introduced novel first-order NPDHG algorithms that overcome the shortcomings of state-of-the-art algorithms for training large-scale, non-smooth Max-Ent models. The crucial ingredient common to our algorithms is the Kullback–Leibler divergence. Using it over the classical proximal operator makes it possible to train large-scale and non-smooth MaxEnt models much more efficiently than the state of the art. In particular, all stepsize parameters and iterations in our algorithms can be calculated on the order $O(mn)$ operations, improving on the complexity bound of $O(\min(m^2n, mn^2))$ operations for computing the optimal stepsize parameters of classical first-order optimization methods, such as the linear PDHG or forward–backward splitting methods. As a consequence, for a given tolerance level $\epsilon > 0$, our algorithms provably compute solutions using on the order of $O(mn/\sqrt{\epsilon})$ or $O(mn/\log(1/\epsilon))$ operations, the order depending on the smoothness of the MaxEnt model and which is optimal with respect to the Nesterov class of optimal first-order methods [39]. Moreover, the computational bottleneck consists of matrix–vector multiplications, which can be trivially parallelized, and so our algorithms exhibit scalable parallelism.

Finally, we have shown that the strong convexity of the Kullback–Leibler divergence with respect to the ℓ_1 norm allows for significantly larger stepsize parameters, thereby speeding up the convergence rate of our algorithms. This was, in particular, observed in Section 5, when we applied our algorithms to fit the WUMI wildfire data set [45] on several non-smooth MaxEnt models to estimate the probabilities of fire occurrences as a function of ecological features. Our algorithms outperformed the state-of-the-art forward–backward splitting and coordinate descent STRUCTMAXENT2 algorithms by at least one order of magnitude. They also yielded results that are in good agreement with physical models of wildfire occurrence [42,43] as well as previous statistical analyses of wildfire drivers [29–31,44]. Future work will explore the scalability of our algorithms for modeling daily scale wildfire probability [80]. It will also explore how the choices of different informative priors affect the MaxEnt models using prediction accuracy as a metric.

We expect our algorithms to provide efficient methods for solving non-smooth Max-Ent models that arise in large-scale machine learning applications beyond the wildfire application explored in this paper. An interesting future direction, albeit tangential to the focus of this work, would be to prove learning guarantees for the MaxEnt problem (3) for arbitrary choices of potential functions, similarly to [19,64], for the choice of ℓ_1 and ℓ_2 norms as potential functions. This would provide some means to assess the "quality" of the probability distribution estimated from the MaxEnt problem. Another direction that would be interesting and impactful is to extend our algorithms to continuous regularized MaxEnt models and prove rigorously that they work: interesting because the continuous version of the MaxEnt problem is an infinite dimensional problem, which makes this problem more technically challenging, impactful because such an algorithm would enable a much broader class of probability distributions to be used in MaxEnt modeling.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Appendix A. Derivation of Update (17)

We divide the derivation into two parts, first deriving an expression for the update p_{k+1} in terms of the previous iterate p_k and then deriving the explicit update for p_{k+1} as in the third line of (16).

Part 1. First, note the minimization problem

$$\min_{p \in \Delta_n} \left\{ \tau_k D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) - \tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \mathbb{E}_p[\boldsymbol{\Phi}] \rangle + D_{\mathrm{KL}}(p \parallel p_k) \right\}$$

has a unique global minimum, denoted by p_{k+1}, in the relative interior of the unit simplex Δ_n for every $k \in \mathbb{N}$. See ([38], Proposition 4.1) for details and note that Proposition 4.1 applies because the Kullback–Leibler divergence is strongly convex with respect to the ℓ_1 norm in its first argument.

Having established that the minimum p_{k+1} exists and is unique, we introduce the Karush–Kuhn–Tucker multipliers $\xi \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{R}^n$ to write the minimization problem for p_{k+1} in terms of an unconstrained optimization problem:

$$\min_{\substack{p \in \mathbb{R}^n \\ \xi \in \mathbb{R} \\ \mu \in \mathbb{R}^n}} \left\{ \tau_k D_{\mathrm{KL}}(p \parallel p_{\mathrm{prior}}) - \tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \mathbb{E}_p[\boldsymbol{\Phi}] \rangle + D_{\mathrm{KL}}(p \parallel p_k) + \xi \left(1 - \sum_{j=1}^n p(j) \right) - \langle \boldsymbol{\mu}, p \rangle \right\}.$$

We can invoke the Karush–Kuhn–Tucker Theorem and use the linearity constraint qualifications to find that p_{k+1} satisfies the first-order optimality condition

$$\tau_k \log(p_{k+1}(j)/p_{\mathrm{prior}}(j)) + \tau_k - \tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \boldsymbol{\Phi}(j) \rangle \\ + \log(p_{k+1}(j)/p_k(j)) + 1 - \xi - [\boldsymbol{\mu}]_j = 0 \tag{A1}$$

for each $j \in \{1, \ldots, n\}$ [81]. Moreover, the complementary slackness condition $\langle \boldsymbol{\mu}, p_{k+1} \rangle = 0$ with the fact that $p_{k+1} > 0$ for every $j \in \{1, \ldots, m\}$ implies $\boldsymbol{\mu} = \mathbf{0}$.

Next, we use the first-order optimality condition to compute $p_{k+1}(j)$ explicitly, rearranging (A1) in terms of $\log(p_{k+1}(j))$, using $\boldsymbol{\mu} = 0$, and taking the exponential yield

$$p_{k+1}(j) = \left(e^{(\xi-1)/(1+\tau_k)-1} \right) \left((p_{\mathrm{prior}}(j))^{\tau_k} p_k(j) e^{\tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \boldsymbol{\Phi}(j) \rangle} \right)^{1/(1+\tau_k)}$$

Since $\sum_{j=1}^n p_{k+1}(j) = 1$, we find that ξ satisfies the relation

$$e^{\tau_k \xi/(1+\tau_k)-1} = \frac{1}{\sum_{j=1}^n \left((p_{\mathrm{prior}}(j))^{\tau_k} p_k(j) e^{\tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \boldsymbol{\Phi}(j) \rangle} \right)^{1/(1+\tau_k)}}.$$

Hence,

$$p_{k+1}(j) = \frac{\left((p_{\mathrm{prior}}(j))^{\tau_k} p_k(j) e^{\tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \boldsymbol{\Phi}(j) \rangle} \right)^{1/(1+\tau_k)}}{\sum_{j=1}^n \left((p_{\mathrm{prior}})^{\tau_k}(j) p_k(j) e^{\tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \boldsymbol{\Phi}(j) \rangle} \right)^{1/(1+\tau_k)}}$$

for each $j \in \{1, \ldots, n\}$. This yields an explicit update for p_{k+1} in terms of the previous iterate p_k.

Part 2. We now induce on k to prove that the explicit expression for the update p_{k+1} in the third line of (16) is correct. For $k = 0$, we have

$$p_0(j) = \frac{p_{\mathrm{prior}}(j) e^{\langle z_0, \boldsymbol{\Phi}(j) \rangle}}{\sum_{j=1}^n p_{\mathrm{prior}}(j) e^{\langle z_0, \boldsymbol{\Phi}(j) \rangle}}$$

and so

$$\left(\left(p_{\text{prior}}(j) \right)^{\tau_0} p_0(j) e^{\tau_0 \langle w_0 + \theta_0(w_0 - w_{-1}), \Phi(j) \rangle} \right)^{1/(1+\tau_0)}$$
$$= \frac{p_{\text{prior}}(j) e^{\langle (z_0 + \tau_0(w_0 + \theta_0(w_0 - w_{-1})))/(1+\tau_0), \Phi(j) \rangle}}{\left(\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_0, \Phi(j) \rangle} \right)^{1/(1+\tau_0)}}.$$

Writing

$$z_1 = (z_0 + \tau_0(w_0 + \theta_0(w_0 - w_{-1})))/(1 + \tau_0),$$

the term inside the exponential on the numerator simplifies to $\langle z_1, \Phi(j) \rangle$. Hence,

$$\left(\left(p_{\text{prior}}(j) \right)^{\tau_0} p_0(j) e^{\tau_0 \langle w_0 + \theta_0(w_0 - w_{-1}), \Phi(j) \rangle} \right)^{1/(1+\tau_0)} = \frac{p_{\text{prior}}(j) e^{\langle z_1, \Phi(j) \rangle}}{\left(\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_0, \Phi(j) \rangle} \right)^{1/(1+\tau_0)}}.$$

It follows on substitution that

$$p_1(j) = \frac{p_{\text{prior}}(j) e^{\langle z_1, \Phi(j) \rangle}}{\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_1, \Phi(j) \rangle}}$$

for every $j \in \{1, \ldots, n\}$.

For arbitrary $k \in \mathbb{N}$, we use the induction hypothesis

$$p_k(j) = \frac{p_{\text{prior}}(j) e^{\langle z_k, \Phi(j) \rangle}}{\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_k, \Phi(j) \rangle}}$$

to find

$$\left(\left(p_{\text{prior}}(j) \right)^{\tau_k} p_k(j) e^{\tau_k \langle w_k + \theta_k(w_k - w_{k-1}), \Phi(j) \rangle} \right)^{1/(1+\tau_k)}$$
$$= \frac{p_{\text{prior}}(j) e^{\langle (z_k + \tau_k(w_k + \theta_k(w_k - w_{k-1})))/(1+\tau_k), \Phi(j) \rangle}}{\left(\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_k, \Phi(j) \rangle} \right)^{1/(1+\tau_k)}}.$$

Writing

$$z_{k+1} = (z_k + \tau_k(w_k + \theta_k(w_k - w_{k-1})))/(1 + \tau_k),$$

the term inside the exponential on the numerator simplifies to $\langle z_{k+1}, \Phi(j) \rangle$. Hence,

$$\left(\left(p_{\text{prior}}(j) \right)^{\tau_k} p_k(j) e^{\tau_{0k} \langle w_k + \theta_k(w_k - w_{k-1}), \Phi(j) \rangle} \right)^{1/(1+\tau_k)} = \frac{p_{\text{prior}}(j) e^{\langle z_{k+1}, \Phi(j) \rangle}}{\left(\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_k, \Phi(j) \rangle} \right)^{1/(1+\tau_k)}}.$$

It follows on substitution that

$$p_{k+1}(j) = \frac{p_{\text{prior}}(j) e^{\langle z_{k+1}, \Phi(j) \rangle}}{\sum_{j=1}^n p_{\text{prior}}(j) e^{\langle z_{k+1}, \Phi(j) \rangle}}$$

for every $j \in \{1, \ldots, n\}$, which proves the desired result.

Appendix B. Summary of Features Extracted from the WUMI Wildfire Data Set

We provide here a summary of all the features used in our statistical analysis. The features are aggregated to a 12 km spatial resolution during data preprocessing. Considering each feature's M antecedent month average and maximum X-day running average components as distinct features, the total number of features considered in our analysis

adds up to 35. We organize the features into five groups with each group reflecting a major environmental driver of fire occurrences. These groups are also an important ingredient of the non-overlapping group lasso MaxEnt model.

Table A1. Summary table for all input features organized by group, identifier, description, spatial resolution of raw data, and source of data.

Feature Group	Identifier	Description	Resolution	Source
Fire weather	VPD	Mean vapor pressure deficit	5 km	Climgrid [82], PRISM [83]
	VPD$^{\max X}$	Maximum X-day running average of VPD; $X \in \{3,7\}$	9 km	UCLA-ERA5 [84]
	Tmax	Daily maximum temperature	5 km	Climgrid
	Tmax$^{\max X}$	Maximum X-day running average of Tmax	9 km	UCLA-ERA5
	Tmin	Daily minimum temperature	5 km	Climgrid
	Tmin$^{\max X}$	Maximum X-day running average of Tmin	9 km	UCLA-ERA5
	Prec	Precipitation total	5 km	Climgrid
	SWE$_{mean}$	Mean snow water equivalent	500 m	NSIDC [85]
	SWE$_{max}$	Daily maximum snow water equivalent	500 m	NSIDC
	FM1000	1000 h dead fuel moisture	4 km	gridMET [86]
	FFWI$^{\max 7}$	Maximum 7-day running average of Fosberg Fire Weather Index	9 km	UCLA-ERA5
	Wind	Mean wind speed	9 km	UCLA-ERA5
	Lightning	Lightning strike density	500 m	NLDN [87,88]
Antecedent	AntVPD$_{M\text{mon}}$	Average VPD in M antecedent months; $M \in \{2,3\}$	5 km	Climgrid
	AntPrec$_{M\text{mon}}$	Average precipitation total in M antecedent months; $M \in \{2,4\}$	5 km	Climgrid
	AntPrec$_{lag1}$	Mean annual precipitation in lag year 1	5 km	Climgrid
	AntPrec$_{lag2}$	Mean annual precipitation in lag year 2	5 km	Climgrid
	AvgSWE$_{3\text{mon}}$	Average snow water equivalent in 3 antecedent months	500 m	NSIDC
Vegetation	Forest	Fraction of forest landcover	30 m	NLCD
	Grassland	Fraction of grassland cover	30 m	NLCD
	Shrubland	Fraction of shrubland cover	30 m	NLCD
	Biomass	Aboveground biomass map	300 m	Ref. [89]
Human	Urban	Fraction of land covered by urban areas	30 m	NLCD
	Camp_num	Mean number of camp grounds	1 km	Open source
	Camp_dist	Mean distance from nearest camp ground	1 km	Open source
	Road_dist	Mean distance from nearest highway	1 km	Open source
	Popdensity	Mean population density	1 km	SILVIS [90]
	Housedensity	Mean housing density	1 km	SILVIS
Topography	Slope	Mean slope	1 m	USGS
	Southness	Mean south-facing degree of slope	1 m	USGS

References

1. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]
2. Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171. [CrossRef]
3. Kapur, J.N. *Maximum-Entropy Models in Science and Engineering*; John Wiley & Sons: Hoboken, NJ, USA, 1989.
4. Gu, Y.; McCallum, A.; Towsley, D. Detecting anomalies in network traffic using maximum entropy estimation. In Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, Berkeley, CA, USA, 19–21 October 2005; p. 32.
5. Bereziński, P.; Jasiul, B.; Szpyrka, M. An entropy-based network anomaly detection method. *Entropy* **2015**, *17*, 2367–2408. [CrossRef]
6. Wainwright, M.J.; Jordan, M.I. *Graphical Models, Exponential Families, and Variational Inference*; Now Publishers Inc.: Hanover, MA, USA, 2008.
7. Berger, A.; Pietra, S.A.D.; Pietra, V.J.D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.
8. Chen, S.F.; Rosenfeld, R. A survey of smoothing techniques for me models. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 37–50. [CrossRef]

9.   Pietra, S.D.; Pietra, V.D.; Lafferty, J. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 380–393.
[CrossRef]

10.   Malouf, R. A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of the COLING-02: The 6th
Conference on Natural Language Learning 2002 (CoNLL-2002), Taipei, Taiwan, 31 August 2002.

11.   Ratnaparkhi, A. *Maximum Entropy Models for Natural Language Processing*; Springer: Boston, MA, USA, 2017; pp. 800–805.

12.   Tsujii, J.; Kazama, J. Evaluation and extension of maximum entropy models with inequality constraints. In Proceedings of the
2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 137–144.

13.   Hayes, B.; Wilson, C. A maximum entropy model of phonotactics and phonotactic learning. *Linguist. Inq.* **2008**, *39*, 379–440.
[CrossRef]

14.   Kwon, T.M. *Tmc Traffic Data Automation for Mndot's Traffic Monitoring Program*; Technical Report; University of Minnesota: Twin
Cities, MN, USA, 2004.

15.   Muttaqin, L.A.; Murti, S.H.; Susilo, B. Maxent (maximum entropy) model for predicting prehistoric cave sites in karst area of
gunung sewu, gunung kidul, yogyakarta. In *Sixth Geoinformation Science Symposium*; SPIE-International Society for Optics and
Photonics: Bellingham, WA , USA, 2019; Volume 11311, p. 113110B.

16.   Granot-Atedgi, E.; Tkačik, G.; Segev, R.; Schneidman, E. Stimulus-dependent maximum entropy models of neural population
codes. *PLoS Comput. Biol.* **2013**, *9*, e1002922. [CrossRef]

17.   Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M.J., II; Bialek, W. The simplest maximum entropy model for collective
behavior in a neural network. *J. Stat. Mech. Theory Exp.* **2013**, *2013*, P03011. [CrossRef]

18.   Savin, C.; Tkačik, G. Maximum entropy models as a tool for building precise neural controls. *Curr. Opin. Neurobiol.* **2017**, *46*,
120–126. [CrossRef] [PubMed]

19.   Dudík, M.; Phillips, S.J.; Schapire, R.E. Performance guarantees for regularized maximum entropy density estimation. In *Learning
Theory*; Shawe-Taylor, J., Singer, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 472–486.

20.   Dudík, M.; Phillips, S.J.; Schapire, R.E. Maximum entropy density estimation with generalized regularization and an application
to species distribution modeling. *J. Mach. Learn. Res.* **2007** 8, 1217–1260.

21.   Elith, J.; Phillips, S.J.; Hastie, T.; Dudík, M.; Chee, Y.E.; Yates, C.J. A statistical explanation of maxent for ecologists. *Divers. Distrib.*
2011, *17*, 43–57. [CrossRef]

22.   Kalinski, C.E. Building Better Species Distribution Models with Machine Learning: Assessing the Role of Covariate Scale and
Tuning in Maxent Models. Ph.D. Thesis, University of Southern California, Los Angeles, CA, USA, 2019.

23.   Merow, C.; Smith, M.J.; Silander, J.A., Jr. A practical guide to maxent for modeling species' distributions: What it does, and why
inputs and settings matter. *Ecography* **2013**, *36*, 1058–1069. [CrossRef]

24.   Phillips, S.J.; Dudík, M.; Schapire, R.E. A maximum entropy approach to species distribution modeling. In Proceedings of the
Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 83.

25.   Phillips, S.J.; Anderson, R.P.; Schapire, R.E. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **2006**,
190, 231–259. [CrossRef]

26.   Phillips, S.J.; Anderson, R.P.; Dudík, M.; Schapire, R.E.; Blair, M.E. Opening the black box: An open-source release of maxent.
Ecography **2017**, *40*, 887–893. [CrossRef]

27.   Schnase, J.L.; Carroll, M.L.; Gill, R.L.; Tamkin, G.S.; Li, J.; Strong, S.L.; Maxwell, T.P.; Aronne, M.E.; Spradlin, C.S. Toward a monte
carlo approach to selecting climate variables in maxent. *PLoS ONE* **2021**, *16*, e0237208. [CrossRef]

28.   Schnase, J.L.; Carroll, M.L. Automatic variable selection in ecological niche modeling: A case study using cassin's sparrow
(peucaea cassinii). *PLoS ONE* **2022**, *17*, e0257502. [CrossRef] [PubMed]

29.   Parisien, M.; Moritz, M.A. Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecol. Monogr.* **2009**, *79*,
127–154. [CrossRef]

30.   Chen, B.; Jin, Y.; Scaduto, E.; Moritz, M.A.; Goulden, M.L.; Randerson, J.T. Climate, fuel, and land use shaped the spatial pattern
of wildfire in California's Sierra Nevada. *J. Geophys. Res. Biogeosci.* **2021**, *126*, e2020JG005786. [CrossRef]

31.   Yu, Y.; Saatchi, S.S.; Walters, B.F.; Ganguly, S.; Li, S.; Hagen, S.; Melendy, L.; Nemani, R.R.; Domke, G.M.; Woodall, C.W. Carbon
Pools across Conus Using the Maxent Model, 2005, 2010, 2015, 2016, and 2017. 2021. Available online: https://daac.ornl.gov/
CMS/guides/CMS_CONUS_Biomass.html (accessed on 1 April 2024).

32.   L'heureux, A.; Grolinger, K.; Elyamany, H.F.; Capretz, M.A.M. Machine learning with big data: Challenges and approaches. *IEEE
Access* **2017**, *5*, 7776–7797. [CrossRef]

33.   Fithian, W.; Hastie, T. Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **2013**, *7*, 1917.
[CrossRef] [PubMed]

34.   Darroch, J.N.; Ratcliff, D. Generalized iterative scaling for log-linear models. In *The Annals of Mathematical Statistics*; Institute of
Mathematical Statistics: Waite Hill, OH, USA, 1972; pp. 1470–1480.

35.   Demchenko, Y.; Grosso, P.; Laat, C.D.; Membrey, P. Addressing big data issues in scientific data infrastructure. In Proceedings
of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013;
pp. 48–55.

36.   Thompson, N.C.; Greenewald, K.; Lee, K.; Manso, G.F. Deep learning's diminishing returns: The cost of improvement is becoming
unsustainable. *IEEE Spectr.* **2021**, *58*, 50–55. [CrossRef]

37. Dyderski, M.K.; Paź, S.; Frelich, L.E.; Jagodziński, A.M. How much does climate change threaten european forest tree species distributions? *Glob. Chang. Biol.* **2018**, *24*, 1150–1163. [CrossRef] [PubMed]

38. Darbon, J.; Langlois, G.P. Accelerated nonlinear primal-dual hybrid gradient algorithms with applications to machine learning. *arXiv* **2021**, arXiv:2109.12222.

39. Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*; Kluwer Academic Publishers: Boston, MA, USA, 2004; Volume 87.

40. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]

41. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.

42. Parisien, M.; Dawe, D.A.; Miller, C.; Stockdale, C.A.; Armitage, O.B. Applications of simulation-based burn probability modelling: A review. *Int. J. Wildland Fire* **2019**, *28*, 913. [CrossRef]

43. Mezuman, K.; Tsigaridis, K.; Faluvegi, G.; Bauer, S.E. The interactive global fire module pyre (v1.0). *Geosci. Model Dev.* **2020**, *13*, 3091–3118. [CrossRef]

44. Buch, J.; Williams, A.P.; Juang, C.S.; Hansen, W.D.; Gentine, P. Smlfire1.0: A stochastic machine learning (sml) model for wildfire activity in the western united states. *Geosci. Model Dev.* **2023**, *16*, 3407–3433. [CrossRef]

45. Juang, C.S.; Williams, A.P.; Abatzoglou, J.T.; Balch, J.K.; Hurteau, M.D.; Moritz, M.A. Rapid growth of large forest fires drives the exponential response of annual forest-fire area to aridity in the western united states. *Geophys. Res. Lett.* **2022**, *49*, e2021GL097131. [CrossRef]

46. Ekeland, I.; Temam, R. *Convex Analysis and Variational Problems*; SIAM: New Delhi, India, 1999.

47. Beck, A.; Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* **2003**, *31*, 167–175. [CrossRef]

48. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.

49. Kemperman, J.H.B. On the optimum rate of transmitting information. In *Probability and Information Theory*; Springer: Berlin/Heidelberg, Germany, 1969; pp. 126–169.

50. Kullback, S. A lower bound for discrimination information in terms of variation (corresp.). *IEEE Trans. Inf. Theory* **1967**, *13*, 126–127. [CrossRef]

51. Pinsker, M.S. *Information and Information Stability of Random Variables and Processes*; Holden-Day: San Francisco, CA, USA, 1964.

52. Altun, Y.; Smola, A. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 139–153.

53. Dewaskar, M.; Tosh, C.; Knoblauch, J.; Dunson, D.B. Robustifying likelihoods by optimistically re-weighting data. *arXiv* **2023**, arXiv:2303.10525.

54. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]

55. Tay, J.K.; Narasimhan, B.; Hastie, T. Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* **2023**, *106*, 1. [CrossRef] [PubMed]

56. Foucart, S.; Rauhut, H. *Sparse Solutions of Underdetermined Systems*; Springer: New York, NY, USA, 2013; pp. 41–59.

57. Guide, M.E.; Jbilou, K.; Koukouvinos, C.; Lappa, A. Comparative study of l 1 regularized logistic regression methods for variable selection. In *Communications in Statistics-Simulation and Computation*; Taylor & Francis, Inc., Philadelphia, PA, USA 2020; pp. 1–16.

58. Zanon, M.; Zambonin, G.; Susto, G.A.; McLoone, S. Sparse logistic regression: Comparison of regularization and bayesian implementations. *Algorithms* **2020**, *13*, 137. [CrossRef]

59. Lee, J.D.; Recht, B.; Srebro, N.; Tropp, J.; Salakhutdinov, R.R. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2010; Volume 23.

60. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Data Mining, Inference, and Prediction; Springer Series in Statistics; Springer: New York, NY, USA, 2009.

61. Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.* **2007**, *1*, 302–332. [CrossRef]

62. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [CrossRef]

63. Hastie, T.; Qian, J.; Tay, K. An Introduction to Glmnet. 2021. Available online: https://glmnet.stanford.edu/articles/glmnet.html (accessed on 17 February 2021).

64. Cortes, C.; Kuznetsov, V.; Mohri, M.; Syed, U. Structural maxent models. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, Lille, France, 6–11 July 2015; ICML'15; pp. 391–399.

65. Chambolle, A.; Bock, T. An introduction to continuous optimization for imaging. *Acta Numer.* **2016**, *25*, 161–319. [CrossRef]

66. Daubechies, I.; Defrise, M.; De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math. J. Issued Courant Inst. Math. Sci.* **2004**, *57*, 1413–1457. [CrossRef]

67. Dudík, M.; Schapire, R.E. Maximum entropy distribution estimation with generalized regularization. In *International Conference on Computational Learning Theory*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 123–138.

68. Mann, G.; McDonald, R.; Mohri, M.; Silberman, N.; Walker, D., IV. Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2009; Volume 22.

69. Andrew, G.; Gao, J. Scalable training of l 1-regularized log-linear models. In Proceedings of the 24th International Conference on Machine Learning, Vienna, Austria, 21–27 July 2007; pp. 33–40.

70. Lee, S.-I.; Lee, H.; Abbeel, P.; Ng, A.Y. Efficient L1 regularized logistic regression. *AAAI* **2006**, *6*, 401–408. .

71. Yuan, G.-X., Chang, K.-W.; Hsieh, C.-J.; Lin, C.-J. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *J. Mach. Learn. Res.* **2010**, *11*, 3183–3234.

72. Yu, H.-F.; Huang, F.-L.; Lin, C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.* **2011**, *85*, 41–75. [CrossRef]

73. Littell, J.S.; McKenzie, D.; Peterson, D.L.; Westerling, A.L. Climate and wildfire area burned in western u.s. ecoprovinces, 1916–2003. *Ecol. Appl.* **2009**, *19*, 1003–1021. [CrossRef]

74. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.

75. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; Chapman & Hall/CRC Texts in Statistical Science; Taylor & Francis: Boca Raton, FL, USA, 2013.

76. Figueiredo, M.A.T.; Nowak, R.D. Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Trans. Image Process.* **2001**, *10*, 1322–1331. [CrossRef] [PubMed]

77. Lions, P.-L.; Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* **1979**, *16*, 964–979. [CrossRef]

78. Hiriart-Urruty, J.-B.; Lemaréchal, C. *Convex Analysis and Minimization Algorithms I: Fundamentals*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1993; Volume 305.

79. Condat, L. Fast projection onto the simplex and the l 1 ball. *Math. Program.* **2016**, *158*, 575–585. [CrossRef]

80. Keeping, T.; Harrison, S.P.; Prentice, I.C. Modelling the daily probability of wildfire occurrence in the contiguous United States. *Environ. Res. Lett.* **2024**, *19*, 024036. [CrossRef]

81. Boyd, S.; Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.

82. Vose, R.; Applequist, S.; Squires, M.; Durre, I.; Menne, M.; Williams, C.; Fenimore, C.; Gleason, K.; Arndt, D. Improved historical Temperature and Precipitation time series for U.S. climate divisions. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 1232–1251. [CrossRef]

83. Daly, C.; Gibson, W.; Doggett, M.; Smith, J.; Taylor, G. Up-to-date monthly climate maps for the conterminous United States. In Proceedings of the 14th AMS Conference on Applied Climatology, 84th AMS Annual Meeting Combined Preprints, Seattle, WA, USA, 11 January 2004.

84. Rahimi, S.; Krantz, W.; Lin, Y.; Bass, B.; Goldenson, N.; Hall, A.; Jebo, Z.; Norris, J. Evaluation of a reanalysis-driven configuration of wrf4 over the western united states from 1980–2020. *J. Geophys. Res. Atmos.* **2022**, *127*, e2021JD035699. [CrossRef]

85. Zeng, X.; Broxton, P.; Dawson, N. Snowpack change from 1982 to 2016 over conterminous United States. *Geophys. Res. Lett.* **2018**, *45*, 12940–12947. [CrossRef]

86. Abatzoglou, J.T. Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.* **2013**, *33*, 121–131. [CrossRef]

87. Wacker, R.S.; Orville, R.E. Changes in measured lightning flash count and return stroke peak current after the 1994 U.S. National Lightning Detection Network upgrade: 1. Observations. *J. Geophys. Res. Atmos.* **1999**, *104*, 2151–2157. [CrossRef]

88. Orville, R.E.; Huffines, G.R. Cloud-to-Ground lightning in the United States: NLDN results in the first decade, 1989–98. *Mon. Weather. Rev.* **2001**, *129*, 1179–1193. [CrossRef]

89. Spawn, S.A.; Sullivan, C.C.; Lark, T.J.; Gibbs, H.K. Harmonized global maps of above and belowground biomass carbon density in the year 2010. *Sci. Data* **2020**, *7*, 112. [CrossRef] [PubMed]

90. Radeloff, V.C.; Hammer, R.B.; Stewart, S.I.; Fried, J.S.; Holcomb, S.S.; McKeefry, J.F. The Wildland-Urban Interface in the United States. *Ecol. Appl.* **2005**, *15*, 799–805. [CrossRef]