Variational Disentanglement for Rare Event Modeling

Zidi Xiu,[1] **Chenyang Tao,** [1]
Michael Gao, [1] **Connor Davis,** [2] **Benjamin A. Goldstein,** [1] **Ricardo Henao** [1]
[1] Duke University
[2] Duke Institute for Health Innovation
zidi.xiu@duke.edu, chenyang.tao@duke.edu, ricardo.henao@duke.edu

Abstract

Combining the increasing availability and abundance of healthcare data and the current advances in machine learning methods have created renewed opportunities to improve clinical decision support systems. However, in healthcare risk prediction applications, the proportion of cases with the condition (label) of interest is often very low relative to the available sample size. Though very prevalent in healthcare, such imbalanced classification settings are also common and challenging in many other scenarios. So motivated, we propose a variational disentanglement approach to semi-parametrically learn from rare events in heavily imbalanced classification problems. Specifically, we leverage the imposed extreme-distribution behavior on a latent space to extract information from low-prevalence events, and develop a robust prediction arm that joins the merits of the generalized additive model and isotonic neural nets. Results on synthetic studies and diverse real-world datasets, including mortality prediction on a COVID-19 cohort, demonstrate that the proposed approach outperforms existing alternatives.

Introduction

Early identification of in-hospital patients who are at imminent risk of life-threatening events, *e.g.*, death, ventilation or intensive care unit (ICU) transfer, is a critical subject in clinical care (Bedoya et al. 2019). Especially during a pandemic like COVID-19, the needs for healthcare change dramatically. With the ability to accurately predict the risk, an automated triage system will be well-positioned to help clinicians better allocate resources and attention to those patients whose adverse outcomes can be averted if early intervention efforts were in place.

Despite the great promise it holds, with the richness of modern Electronic Health Record (EHR) repositories, the construction of such a system faces practical challenges. A major obstacle is the scarcity of patients experiencing adverse outcomes of interest. In the COVID-19 scenario, which we consider in our experiments, the mortality of patients tested positive at the Duke University Health System (DUHS) is slightly lower than 3%. Further, in another typical EHR dataset we consider, less than 5% of patients are reported to suffer adverse outcomes (ICU transfer or death).

In these *low-prevalence* scenarios, commonly seen in clinical practice, standard classification models such as logistic regression suffer from *majority domination*, in which models tend to favor the prediction accuracy of majority groups. This is clearly undesirable for critical-care applications, given the high false negative rates (Type-II error), in which patients in urgent need of care could be falsely categorized.

Situations where the distribution of labels is highly skewed and the accuracy of the minority class bears particular significance (Dal Pozzolo et al. 2017; Lu, Guo, and Li 2020; Machado and Lopes 2020) have been associated with the name *imbalanced dataset* (He and Garcia 2009), whereas the methods dealing with such cases are coined *extreme classification* (Zong, Huang, and Chen 2013). Under such a setting, the lack of representation of minority cases severely undermines the ability of a standard learner to discriminate, relative to balanced datasets (Mitchell 1999). Consequently, these solutions do not generalize well on minority classes, where the primary interest is usually focused.

To address such a dilemma, several remedies have been proposed to account for the imbalance between class representations. One of the most popular strategies is the sampling-based adjustment, where during training, a model oversamples the minority classes (or undersamples the majority classes) to create balance artificially (Holte et al. 1989; Drummond, Holte et al. 2003). To overcome the biases and the lack of information that naive sampling adjustments might induce, variants have been proposed to maximally preserve the clustering structure of the original dataset (Mani and Zhang 2003; Yen and Lee 2009) and to promote diversity of oversampling schemes (Han, Wang, and Mao 2005). Alternatively, cost-sensitive weighting where minority losses are assigned larger weights provides another popular option via tuning the relative importance of minority classes (Elkan 2001; Munro et al. 1996; Zhou and Liu 2005).

While the above two strategies introduce heuristics to alleviate the issues caused by class imbalance, importance sampling (IS) offers a principled treatment that flexibly combines the merits of the two (Hahn and Jeruchim 1987; Heidelberger 1995). Each example is sampled with the probability of a pre-specified importance weight, and with the weight's inverse when accounting for the relative contribution in the overall loss. This helps to flexibly tune the

representation of rare events during training, without biasing the data distribution (Heidelberger 1995; Shimodaira 2000; Gretton et al. 2009). It is important to note that, poor choice of importance weights may result in uncontrolled variance that destabilizes training (Robert and Casella 2013; Botev and Kroese 2008), calling for adaptive (Rubinstein and Kroese 2013) or variance reduction schemes (Rubinstein and Kroese 2016) to protect against such degeneracy.

Apart from the above strategies that fall within the standard empirical risk minimization framework, recent developments explicitly seek better generalization for the minority classes. One such example is the *one-class classification* that aims to capture one target class from a general population (Tax 2002). *Meta-learning* and *few-shot learning* strategies instead trying to transfer the knowledge learned from data-rich classes to facilitate the learning of data-scarce classes (Böhning, Mylona, and Kimber 2015; Finn, Abbeel, and Levine 2017). Additionally, non-cross-entropy based losses or penalties have been proved useful to imbalanced classification tasks (Weinberger and Saul 2009; Huang et al. 2016; Oh Song et al. 2016). For instance, the Focal loss (Lin et al. 2017) up-weights the harder examples, and Cao et al. (2019) introduced a label-distribution-aware margin loss encouraging minority classes having larger margins.

In this work, we present a novel solution called *variational inference for extremals* (VIE), capitalizing on the learning of more generalizable representations for the minority classes. Our proposal is motivated by the observation that the statistical features of "rarity" have been largely overlooked in the current literature of rare-event modeling. And the uncertainties of rare-events are often not considered. Framed under the Variational Inference framework, we formulate our model with the assumption that the extreme presentation of (unobserved) latent variables can lead to the occurrence (or the inhibition) of rare events. This encourages the accurate characterization of the tail distribution of the data representation, which has been missed by prior work to the best of our knowledge. Building upon the state-of-the-art machine learning techniques, our solution features the following contributions: (i) the model accounts for representation uncertainty based on variational inference; (ii) the adoption of mixed Generalized Pareto priors to promote the learning of heavy-tailed feature representations; and (iii) integration of additive isotonic regression to disentangle representation and facilitate generalization. We demonstrate how our framework facilitates both model generalization and interpretation, with strong empirical performance reported across a wide-range of benchmarks.

Background

To simplify our presentation, we focus on the problem of rare event classification for binary outcomes. The generalization to the multiple-class scenario is simple and presented in the Supplementary Material (SM)[1]. Let $D = \{x_i, y_i\}_{i=1}^N$ be a dataset of interest, where x_i and y_i denote predictors and outcomes, respectively, and N is the sample size. Without loss of generality, we denote $y = 1$ as the minority event label (indicating the occurrence of an event of interest), and $y = 0$ as the majority label.

In the following, we will briefly review the three main techniques we used in this work, namely, *variational inference* (VI), *extreme value theory* (EVT), and *additive isotonic regression*. VI allows for approximate maximum likelihood inference while accounting for data uncertainty. EVT provides a principled and efficient way to model extreme, heavy-tailed representations. Additive isotonic regression further introduces monotonic constraints to *disentangle* the contribution of each latent dimension to the outcome.

Variational Inference

Consider a latent variable model $p_\theta(v, z) = p_\theta(v|z)p(z)$, where $v \in \mathbb{R}^m$ is the observable data, $z \in \mathbb{R}^p$ is the unobservable latent variable, and θ represents the parameters of the likelihood model, $p_\theta(v|z)$. The marginal likelihood $p_\theta(v) = \int p_\theta(v, z)\mathrm{d}z$ requires integrating out the latent z, which typically, for complex distributions, does not enjoy a closed-form expression. This intractability prevents direct maximum likelihood estimation for θ in the latent variable setup. To overcome this difficulty, Variational Inference (VI) optimizes computationally tractable variational bounds to the marginal log-likelihood (Kingma and Welling 2014; Chen et al. 2018). Concretely, the most popular choice of VI optimizes the following Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}(v; p_\theta(v, z), q_\phi(z|v)) &\triangleq \mathbb{E}_{Z\sim q_\phi(z|v)}\left[\log \frac{p_\theta(v, Z)}{q_\phi(Z|v)}\right] \\ &\leq \log p_\theta(v), \qquad (1)\end{aligned}$$

where $q_\phi(z|v)$ is an approximation to the true (unknown) posterior $p_\theta(z|v)$, and the inequality is a direct result of Jensen's inequality. The variational gap between the ELBO and true marginal log-likelihood, *i.e.*, $\log p_\theta(v) - \mathrm{ELBO}(v; p_\theta(v, z), q_\phi(z|v))$, is given by the Kullback–Leibler (KL) divergence between posteriors, *i.e.*, $\mathrm{KL}(q_\phi(z|v)||p_\theta(z|v)) = \mathbb{E}_{Z\sim q_\phi(z|v)}[\log q_\phi(Z|v)] - \mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(Z|v)]$, which implies that the ELBO tightens as $q_\phi(z|v)$ approaching the true posterior $p_\theta(z|v)$. For estimation, we seek parameters θ and ϕ that maximize the ELBO in (1).

Given a set of observations $\{v_i\}_{i=1}^N$ sampled from data distribution $v \sim p_d(v)$, maximizing the expected ELBO is also equivalent to minimizing the KL divergence $\mathrm{KL}(p_d(v) \parallel p_\theta(v))$ between the empirical and model distributions. When $p_\theta(v|z)$ and $q_\phi(z|v)$ are specified as neural networks, the resulting architecture is commonly known as the *variational auto-encoder* (VAE) (Kingma and Welling 2014), where $q_\phi(z|v)$ and $p_\theta(v|z)$ and are known as *encoder* and *decoder*, respectively. Note that $q_\phi(z|v)$ is often used for subsequent inference tasks on new data.

Extreme Value Theory

Extreme Value Theory (EVT) provides a principled probabilistic framework for describing events with extremely low probabilities, which we seek to exploit for better rare event modeling. In particular, we focus on the *exceedance* models, where we aim to capture the asymptotic statistical behavior of values surpassing an extreme threshold (Davison

[1]The SM can be found at https://arxiv.org/abs/2009.08541

and Smith 1990; Tao et al. 2017), which we briefly review below following the notation of Coles et al. (2001). Without loss of generality, we consider exceedance to the right, *i.e.*, values greater than a threshold u. For a random variable X, the conditional cumulative distribution of exceedance level x beyond u is given by $F_u(x) = P(X - u \leq x | X > u) = \frac{F(x+u)-F(u)}{1-F(u)}$, where $x > 0$ and $F(x)$ denotes the cumulative density function for X.

A major result from EVT is that under some mild regularity conditions, *e.g.*, continuity at the right end of $F(x)$ and others, $F_u(x)$ will converge to the family of *Generalized Pareto Distributions* (GPD) regardless of $F(x)$, as u approaches the right support boundary of $F(x)$ (Balkema and De Haan 1974; Pickands III et al. 1975), *i.e.*, $\lim_{u\to\infty} F_u(x) \xrightarrow{L_\infty} G_{\xi,\sigma,u}(x)$ (Falk, Hüsler, and Reiss 2010), where $\text{GPD}_{\xi,\sigma,u}(x)$ is of the form

$$G_{\xi,\sigma,u}(x) = \begin{cases} 1 - [1 + \xi(x-u)/\sigma]^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ 1 - \exp[-(x-u)/\sigma], & \text{if } \xi = 0 \end{cases} \quad (2)$$

where σ is a positive scale parameter. When $\xi < 0$ the exceedance x has bounded support $u \leq x \leq u - \sigma/\xi$, otherwise when $\xi \geq 0$, x ia unbounded on the right. A major implication of this asymptotic behavior is that, for modeling extreme values, one only needs to fit extreme samples to the log-likelihood function of the GPD.

Additive Isotonic Regression

Also known as monotonic regression, isotonic regression is a non-parametric regression model that constrains the relation between predictor and outcome to be monotonic, (*e.g.*, non-decreasing $f(a) \leq f(b)$ for $a \leq b$) (Barlow et al. 1972). Such monotonic constraint is a natural and flexible extension to the standard linear relation assumed by many statistical models. To accommodate multi-covariate predictors, additive isotonic regression combines isotonic models for each individual one-dimensional predictor (Bacchetti 1989). Standard implementations often involve specialized algorithms, such as local scoring algorithms (Hastie 2017) and the alternating conditional expectation (ACE) method of Breiman and Friedman (1985). All these approaches typically require costly iterative computations and are not scalable to large datasets. Here we consider recent advances in unconstrained monotonic neural networks, which allow for efficient and flexible end-to-end learning of monotonic relations with robust neural nets based on standard training schemes such as stochastic gradient descent (Sill 1998; Wehenkel and Louppe 2019).

Variational Inference OF Extremals

The proposed model is based on the hypothesis that *extreme events are driven by the extreme values of some latent factors*. Specifically, we propose to recast the learning of low-prevalence events into the learning of extreme latent representations, thus amortizing the difficulties associated with directly modeling rare events as outcomes. To allow for more efficient learning from the rare events, we make some further assumptions to regularize the latent representation: (i) *effect disentanglement*: the contribution from each dimension of the latent representation to the event occurrence is additive; (ii) *effect monotonicity*: there is a monotonic relation between the outcome likelihood and the values of each dimension of the latent representation. The key to the proposed approach is using an additive isotonic neural network to model the one-dimensional disentangled monotonic relations from a latent representation, which is obtained via variational inference. Specifically, we impose an EVT prior to explicitly capture the information from the few minority group samples into the tail behavior of the extreme representation. Below we provide the rationale for our choices followed by a description of all model components.

Disentanglement & additive isotonic regression. Consistent with assumptions (i) and (ii), we posit a scenario in which the underlying representation of extreme events is more frequent at the far end of the representation spectrum, for which additive isotonic regression is ideal. The disentanglement consists of modeling each latent dimension individually, thus avoiding the curse of dimensionality when modeling combinatorial effects with few examples. Further, the monotonicity constraint imposed by the isotonic regression model restricts possible effect relations, thereby improving generalization error by learning with a smaller, yet still sufficiently expressive, class of models (Bacchetti 1989).

EVT & VI. Note that the spread of representation of extreme events is expected to be more uncertain relative to those of the normal, more abundant events, due to a few plausible causes: (i) extreme events represent the breakdown of system normality and are expected to behave in uncertain ways; (ii) there is only a small number of examples available for the extreme events, so the learned feature encoder will tend to be unreliable. As a result, we can safely assume that the encoded features associated with the extremes events will lie outside the effective support of the Gaussian distribution assumed by the standard VI model. In other words, the representation of the events can manifest as a heavy-tailed distribution. This will compromise the validity and generalizability of a prediction model if not dealt with appropriately. So motivated, we explicitly model the distribution of the extreme underlying representations via EVT. Using EVT, we decouple the learning of the tail end of the representation distribution. Since EVT-based estimation only requires very few parameters, it allows for accurate modeling with a small set of tail-end samples. Further, in combination with the variational inference framework, it accounts for representation uncertainty via the use of a stochastic encoder, which further strengthens model robustness.

Benefits of heavy-tailed modeling. A few other considerations further justify modeling with a heavy-tailed distribution for the extreme event representation. One obvious benefit is that it allows better model resolution along the representation axis, *i.e.*, better risk stratification. For light-tail representations, extreme examples are clustered in a narrow region where the tail vanishes, thus a standard (light-tailed) learning model will report the average risk in that region. However, if the representations are more spread out, then there is a more gradual change in risk (Figure S1 in the

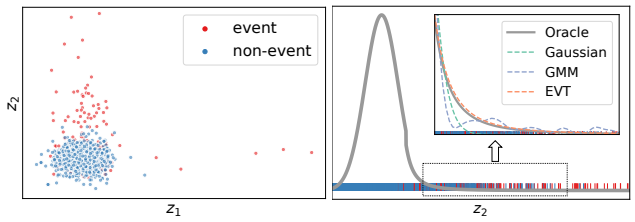

Figure 1: Left: Distribution of a two-dimensional latent space z where the long tail associates with higher risk. Right: Tail estimations with different schemes for the long-tailed data in one-dimensional space. EVT provides more accurate characterization comparing to other mechanisms.

SM), which can be better captured, as shown in Figure 1. Another argument for favoring heavy-tailed representations is that heavy-tailed phenomena are very common in nature (Bryson 1974), and these tail samples are often encoded less robustly due to the lack of training examples. Allowing long-tail representations relieves the burden of an encoder.

Model structure. We consider latent variable model $p_\theta(y, x, z) = p_\theta(y|z)p_\theta(x|z)p(z)$, where $v = \{x, y\}$ are the observed variables. Under the VI framework, similar to (1) we write the ELBO$(v; p_\theta(v, z), q_\phi(z|v))$ as

$$\mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(y|Z)] + \mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(x|Z)] - \mathrm{KL}(q_\phi(z|v) \parallel p(z)) \quad (3)$$

where $p_\theta(y|z)$ is specified as an additive isotonic regression model, $p(z)$ is modeled with Gaussian GPD mixture, and the approximate posterior, $q_\phi(z|v)$, is specified as an inverse auto-regressive flow. Note that unlike in the standard ELBO in (3), we want to drop the term $\mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(x|Z)]$ because we are not interested in modeling the covariates. Note this coincides with the *variational information bottleneck* (VIB) formulation (Alemi et al. 2016). Additionally, the posterior $q_\phi(z|v)$ will not be conditioned on y, but only on x, because in practice, the labels y are not available at inference time. Specifically, we rewrite the objective in (3) as

$$\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) = \mathbb{E}_{Z\sim q_\phi(z|x)}[\log p_\theta(y|Z)] - \beta\mathrm{KL}(q_\phi(z|x) \parallel p(z)), \quad (4)$$

where β is a hyperparameter controlling the relative contribution of the KL term to the objective. Below we provide details for each component of the proposed approach.

Decoder: Additive Monotonic Neural Network

First, let us consider the following monotone mapping $\int_l^z h(s;\theta)ds + \gamma$, consisting on integrating a non-negative function $h(s;\theta)$ specified as a neural network with one-dimensional input, s, and parameterized by θ. The choice of the lower end l is arbitrary, and γ is a bias term. For multi-dimensional latent representation $z \in \mathbb{R}^p$, we write the additive monotonic neural network (AMNN) as

$$H(z;\theta) = \sum_j^p [\alpha_j \int_l^{z_j} h_j(s;\theta)ds] + \gamma, \quad (5)$$

where α_j serves as a weight which controlling the effect directions. In other words, when $\alpha_j > 0$, it can be interpreted as an event stimulator; otherwise it is an event blocker. To ensure $h(s;\theta)$ is non-negative, we apply exponential activation function to the network's output. The integration of z is conducted with numerical integration by the Riemann-Stieltjes method (Davis and Rabinowitz 2007).

Based on (5), we have $\log p_\theta(y|z) = \ell_{\mathrm{CLL}}(y, H(z;\theta))$, where $\ell_{\mathrm{CLL}}(y, a) = \log\{\mathbb{1}_{y=1}(y)(1 - \exp(-\exp(a))) + \mathbb{1}_{y=0}(y)\exp(-\exp(a))\}$ is the complementary log-log (CLL) link, where $\mathbb{1}_{(\cdot)}$ is the indicator function. We prefer CLL over the standard logistic link since the CLL link is more sensitive at the tail end (Aranda-Ordaz 1981).

Latent Extreme Prior: Gaussian GPD Mixture

To better capture the tail behavior of the latent representation, we assume random variable $Z \sim p(z)$ is a mixture of a standard Gaussian distribution truncated at u and a GPD for modeling the tail end thresholded at u, *i.e.*, $F(z) = \Phi(z)$ when $z \le u$ and $F(z) = \Phi(u) + (1 - \Phi(u))G_{\xi,\sigma}(z - u)$ when $z > u$, where $\Phi(z)$ denotes the CDF of a standard Gaussian distribution. Note that for $z > u$, $F(z)$ can be expressed as a GPD with parameters $(\tilde{\xi}, \tilde{\sigma}, \tilde{u})$ (McNeil 1997), where $\tilde{\xi} = \xi$ and if $\xi \neq 0$, $\tilde{\sigma} = \sigma(1 - \Phi(u))^\xi$ and $\tilde{u} = u - \tilde{\sigma}((1 - \Phi(u))^{-\xi} - 1)/\xi$. Otherwise, when $\xi = 0$, $\tilde{\sigma} = \sigma$ and $\tilde{u} = u + \tilde{\sigma}\log(1 - \Phi(u))$. Consequently, the CDF for the mixed GPD is given by

$$F(z) = \mathbb{1}_{(-\infty,u]}(z)\Phi(z) + \mathbb{1}_{(u,\infty)}(z)G_{\tilde{u},\xi,\tilde{\sigma}}(z). \quad (6)$$

For simplicity, we denote the set of parameters in GPD as $\psi = \{\xi_{\mathrm{GPD}}, \sigma_{\mathrm{GPD}}\}$ and the threshold u is a user-defined parameter. In the experiments we set u to $\Phi^{-1}(0.99)$.

Latent Posterior: Inverse Autoregressive Flow

Considering we have adopted a long-tailed GPD prior, we seek a posterior approximation $q_\phi(z|x)$ that is: (i) a flexible parameterization to approximate arbitrary distributions; and (ii) with a tractable likelihood to be able to evaluate the $\mathrm{KL}(q_\phi(z|x) \parallel p(z))$ exactly. We need (i) because the true posterior is likely to exhibit heavy-tailed behavior due to the extended coverage of the GPD prior, and (ii) is to ensure accurate and low-variance Monte Carlo estimation of the KL-divergence at the tail end of the prior. These requirements invalidate some popular choices, *e.g.*, a standard Gaussian posterior is light-tailed, and the implicit neural-sampler-based posterior typical in the work of adversarial variational Bayes (Mescheder, Nowozin, and Geiger 2017), does not have a tractable likelihood.

One model family satisfying the above two requirements is known as the generative flow (Rezende and Mohamed 2015), where simple invertible transformations with tractable log Jacobian determinants are stacked together, transforming a simple base distribution into a complex one, while still having closed-form expressions for the likelihood. In this work, we consider the *inverse autoregressive flow* (IAF) model (Kingma et al. 2016). The flow chain is built as

$$z_t = \mu_t + \sigma_t \odot z_{t-1}, \text{ for } 1 \le t \le T, \quad (7)$$

where $\mu_t \in \mathbb{R}^p$ and $\sigma_t \in \mathbb{R}^p$ are learnable parameters, $\odot$ denotes the element-wise product, z_0 is typically drawn from a p-dimensional Gaussian distribution, $z_0 \sim \mathcal{N}(\mu_0, \text{Diag}(\sigma_0^2))$ where μ_0 and σ_0 are obtained from an initial encoder defined by a neural network given input x with parameter ϕ. A sample from the posterior $q_\phi(z|x)$ is given by z_T, obtained by "flowing" z_0 through (7). Provided the Jacobians $\frac{d\mu_t}{dz_{t-1}}$ and $\frac{d\sigma_t}{dz_{t-1}}$ are strictly upper triangular (Papamakarios, Pavlakou, and Murray 2017), we obtain the following closed-form expression for the log posterior

$$\begin{aligned}\log q(z|x) &= \log q(z_0|x) - \sum_{t=1}^{T} \log\det\left|\frac{dz_t}{dz_{t-1}}\right| \qquad (8)\\ &= -\sum_{j=1}^{p}\left(\frac{1}{2}e_j^2 + \frac{1}{2}\log(2\pi) + \sum_{t=0}^{T}\log\sigma_{t,j}\right),\end{aligned}$$

where $e_j = (x_j - \mu_{0,j})/\sigma_{0,j}$ for the jth dimension.

Posterior Match with Fenchel Mini-Max Learning

We consider an additional modification that explicitly encourages the match of the aggregated posterior $q_\phi(z) = \int q_\phi(z|x)p_d(x)\mathrm{d}x$ to the prior $p(z)$, which has been reported to be vastly successful at improving VAE learning (Mescheder, Nowozin, and Geiger 2017). In our case, $q_\phi(z)$ does not have a closed-form expression for the likelihood ratio of the KL formulation, which motivates us to use a sample-based estimator. We consider the mini-max KL estimator based on the Fenchel duality (Tao et al. 2019; Dai et al. 2018). Concretely, recall the KL can be expressed in its Fenchel dual form[2]

$$\begin{aligned}&\Gamma(p, q_\phi, \nu) = \mathbb{E}_{Z\sim q_\phi(z)}[\nu(Z)] - \mathbb{E}_{Z'\sim p(z)}[\exp(\nu(Z'))]\\ &\text{KL}(q_\phi(z) \parallel p(z)) = \max_{\nu\in\mathcal{F}} \Gamma(p, q, \nu), \qquad (9)\end{aligned}$$

where $\nu(z)$ is commonly known as the critic function in the adversarial learning literature, and we maximize wrt $\nu(z)$ in the space of all functions $\mathcal{F}$, modeled with a deep neural network. We use (4) and (9) to derive an augmented ELBO that further penalizes the discrepancy between the aggregated posterior and the prior, *i.e.*, $\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) - \lambda\text{KL}(q_\phi(z) \parallel p(z))$, where λ is a regularization hyperparameter (Chen, Feng, and Lu 2018). Solving for this objective results in the following mini-max game

$$\max_{\theta,\phi}\min_{\nu} \Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) - \lambda\Gamma(p_\theta, q_\phi, \nu), \quad (10)$$

where β and λ are regularization hyperparameters. In a similar vein to β-VAE and adversarial variational Bayes (AVB), our objective leverages β, $\lambda > 0$ to balance the prediction accuracy and the complexity of the latent representation via KL regularization. Further, from $\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x))$ in (4), note that the decoder $p_\theta(y|z)$ is obtained from the additive neural network in (5), $p_\psi(z)$ is the Gaussian GPD mixture with CDF in (6), $q_\phi(z|x)$ is the autoregressive flow implied by (7) and $\nu(z;\omega)$ is the critic function specified as a neural network and parameterized by ω.

[2]We have removed the constant term for notational clarity.

Algorithm 1: Variational Inference with Extremals.

Data: $\mathcal{D} = (x, y)$. x: inputs, y: labels
Networks and parameters:
`Init-Encoder`$(x, \epsilon; \phi)$: Initial encoder network;
`IAF`$(z; \phi)$: recursive autoregressive neural network;
$\nu(z; \omega)$: critic neural network;
`AMNN`$(z; \theta)$: additive monotonic neural net;
prior: $p_\psi(z)$ =`MixedGPD`$(z;\psi,u)$, ψ=$\{\xi_{\text{GPD}},\sigma_{\text{GPD}}\}$
Initialize: `Init-Encoder`, `IAF`, ν, `AMNN`, ψ
for *iteration* $k \in \{1, \ldots, K\}$ **do**
  Sample $\{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{D}$, $\{\epsilon_i\}_{i=1}^m$ from $p(\epsilon)$
  $[\mu_0, \sigma_0]$ =`Init-Encoder`$(x, \epsilon; \phi)$
  Sample z_{pr} from $p_\psi(z)$, $\boldsymbol{z}_0$ from $\mathcal{N}(\mu_0, \Sigma_0)$
  Compute $l_{\texttt{post}} := \log q_\phi(z_0|x)$
  for *step* $t \in \{1, \ldots, T\}$ **do**
    $[\mu_t, \sigma_t]$ =`IAF`$(z_{t-1}; \phi)$, $z_t = \mu_t + \sigma_t \odot z_{t-1}$
    $l_{\texttt{post}} = l_{\texttt{post}} - \sum(\log\sigma_t)$
  end
  $\log p_\theta(y|z_T) = \ell_{\text{CLL}}(y, \text{AMNN}(z_T; \theta))$
  Descend ω by $\nabla_\omega \frac{1}{m}\sum[\nu_\omega(z_{\text{pr}}) - \log\nu_\omega(z_T)]$
  Ascend $\Omega = \{\phi, \psi, \theta\}$ by
  $\nabla_\Omega \frac{1}{m}\sum[\log p_\theta(y|z_T) - \log\nu_\omega(z_T) - \text{KL}]$,
   where $\text{KL} = l_{\texttt{post}} - \log p_\psi(z_T)$
end

To avoid collapsing to suboptimal local minima, we train the encoder arm more frequently to compensate for the detrimental posterior lagging phenomenon (He et al. 2019). The pseudo-code for the proposed VIE is summarized in Algorithm 1 and detailed architecture can be found in the SM.

Related Work

Rare-event modeling with regression. Initiated by King and Zeng (2001), the discussion on how to handle the unique challenges presented by rare-event data for regression models has attracted extensive research attention. The statistical literature has mainly focused on bias correction for sampling (Fithian and Hastie 2014) and estimation (Firth 1993), driven by theoretical considerations in maximum likelihood estimation. However, their assumptions are often violated in the face of modern datasets (Sur and Candès 2019), characterized by high-dimensionality and complex interactions. Our proposal approaches a solution from a representation learning perspective (Bengio, Courville, and Vincent 2013), by explicitly exploiting the statistical regularities of extreme values to better capture extreme representations associated with rare events.

Re-sampling and loss correction. Applying statistical adjustments during model training is a straightforward solution to re-establish balance, but often associated with obvious caveats. For example, the popular down-sampling and up-sampling (He and Garcia 2009) discard useful information or introduce artificial bias, exacerbating the chances of capturing spurious features that may harm generalization (Drummond, Holte et al. 2003; Cao et al. 2019), and their performance gains may be limited (Byrd and Lipton

					Average AUC (standard deviation)		Average AUPRC (standard deviation)	
	Prior	Encoder	Decoder	Prior Match	n=5k	n=20k	n=5k	n=20k
VAE	Gaussian	Gaussian	MLP	True	0.552 (0.092)	0.674 (0.020)	0.026 (0.010)	0.061 (0.017)
VAE-GPD	mixed GPD	Gaussian	AMNN	False	0.569 (0.062)	0.653 (0.027)	0.021 (0.003)	0.035 (0.013)
IAF-GPD	mixed GPD	IAF	AMNN	False	0.511(0.021)	0.665 (0.029)	0.017(0.002)	0.025 (0.008)
Fenchel-GPD	mixed GPD	Implicit	AMNN	True	0.623 (0.036)	0.694 (0.021)	0.037 (0.010)	0.062 (0.026)
VIE	mixed GPD	IAF	AMNN	True	**0.684** (0.031)	**0.701** (0.017)	**0.050** (0.009)	**0.079** (0.025)
Oracle (with 90% confidence interval)					0.704 [0.662, 0.751]		0.092 [0.058, 0.141]	

Table 1: Ablation study of VIE with different combinations of architectures on realistic synthetic datasets with 1% event rate. The oracle model has used the ground-truth model parameters to predict.

2019). While traditionally tuned by trial and error, recent works have explored automated weight adjustments (Lin et al. 2017; Zhang et al. 2020) , and principled loss correction that factored in class-size differences (Cui et al. 2019). Our contribution is orthogonal to these developments and promises additional gains when used in synergy.

Transferring knowledge from the majority classes. Adapting the knowledge learned from data-rich classes to their under-represented counterparts has shown success in few-shot learning, especially in the visual recognition field (Wang, Ramanan, and Hebert 2017; Chen et al. 2020) , and also in the clinical settings (Böhning, Mylona, and Kimber 2015). However, their success often critically depends on strong assumptions, the violation of which typically severely undermines performance (Wang et al. 2020). Related are the one-class classification (OCC) models (Tax 2002), assuming stable patterns for the majority over the minority classes. Our assumptions are weaker than those made in these model categories, and empirical results also suggest the proposed VIE works more favorably in practice (see experiments).

Experiments

We carefully evaluate the proposed VIE on a diverse set of realistic synthetic data and real-world datasets with different degrees of imbalance. Our implementation is based on PyTorch, and code to replicate our experiments are available from https://github.com/ZidiXiu/VIE/. We provide additional experiments and analyses in the SM.

Baseline Models We consider the following set of competing baselines to compare the proposed solution: LASSO regression (Tibshirani 1996), MLP with re-sampling and re-weighting (MLP), Importance-Weighting model (IW) (Byrd and Lipton 2019), FOCAL loss (Lin et al. 2017), Label-Distribution-Aware Margin loss (LDAM) (Cao et al. 2019), and SVD based one-class classification model (Deep-SVDD) (Ruff et al. 2018). We evaluate baseline models performance on test set with the best performed hyper-parameters on validation dataset. For detailed settings please refer to the SM.

Evaluation Metrics To quantify model performance, we consider AUC and AUPRC. AUC is the area under the Receiver Operating Characteristic (ROC) curve, which provides a threshold-free evaluation metric for classification model performance. AUC summarizes the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). AUPRC evaluates the area under Precision-Recall (PR) curve and summarizes the trade-off between TPR and True Predictive Rate. We discuss other metrics in the SM. In simulation studies, we repeat simulation ten times to obtain empirical AUC and AUPRC confidence intervals. For real world datasets, we applied bootstrapping to estimate the confidence intervals.

Ablation Study for VIE

VIE applies a few state-of-art techniques in variational inference in order to achieve optimal performance. In this section, we decouple their contributions via an ablation study, to justify the necessity of including those techniques in our final model. To this end, we synthesize a semi-synthetic dataset based on the Framingham study (Mitchell et al. 2010), a long-term cardiovascular survival cohort study. We use a realistic model to synthesize data from the real-world covariates under varying conditions, *i.e.*, different event rates, sample size, non-linearity, *etc.* More specifically, we use the CoxPH-Weibull model (Bender, Augustin, and Blettner 2005) to simulate the survival times of patients $T = \{\frac{-\log U}{\lambda \exp(g(x))}\}^{1/\nu}$, where $g(x)$ is either a linear function or a randomly initialized neural net. Our goal is to predict whether the subject will decease within a pre-specified time frame, *i.e.*, $T < t_0$. Via adjusting the cut-off threshold t_0, we can simulate different event rates. A detailed description of the simulation strategy is in the SM.

We experiment with different combinations of advanced VI techniques, as summarized in Table 1. Limited by space, we report results at 1% event rate with $g(\cdot)$ set to a randomly initialized neural network under various sample sizes. Additional results on linear models and other synthetic datasets are consistent and can be found in the SM. IAF and GPD only variants perform poorly, even compared to the vanilla VAE solution. This is possibly due to the fact that priors are mismatched. Explicitly matching to the prior via Fenchel mini-max learning technique improves performance. However, without using an encoder with a tractable likelihood, the model cannot directly leverage knowledge from the GPD prior likelihood. Stacked together (mixed GPD+IAF+Fenchel), our full proposal of VIE consistently outperforms its variants, approaching oracle performance in the large sample regime.

	Average AUC						Average AUPRC					
	COVID		InP		SEER		COVID		InP		SEER	
Event category	Mortality	Combined	12h	168h	3mo	11mo	Mortality	Combined	12h	168h	3mo	11mo
LASSO	0.856	0.853	0.822	0.760	0.888	0.845	0.235	**0.542**	0.092	0.216	0.140	0.309
MLP	0.862	0.854	0.824	0.768	0.885	0.856	0.225	0.531	0.093	0.221	0.169	0.322
DeepSVDD	NA	NA	0.633	0.551	0.592	0.572	NA	NA	0.020	0.063	0.026	0.068
IW	0.856	0.860	0.776	0.728	0.798	0.832	0.193	0.511	0.073	0.165	0.123	0.274
Focal	0.829	0.854	0.750	0.705	0.868	0.835	0.238	0.484	0.044	0.149	0.141	0.263
LDAM	0.857	0.843	0.819	0.774	0.893	0.755	0.202	0.535	0.086	0.197	0.177	0.332
VIE	**0.883**	**0.867**	**0.840**	**0.780**	**0.895**	**0.862**	**0.268**	0.535	**0.100**	**0.240**	**0.189**	**0.345**

Table 2: Average AUC and AUPRC from real-world datasets.

	COVID	INP	SEER
sample size	25,315	67,655	68,082
dimension	1268(668)	73(39)	789(771)
event rate (%)	2.6%, 8%	$1 \sim 5\%$	$1 \sim 5\%$

Table 3: Summary statistics for real-world datasets.

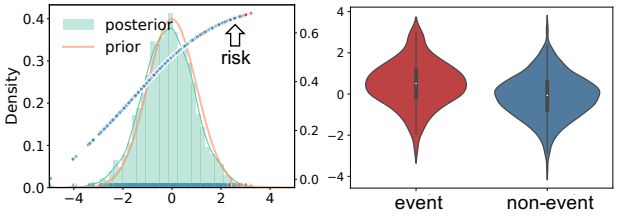

Figure 2: First latent dimension from the InP dataset (1% event rate). Left: Learned prior and posterior distribution, and monotonic predicted risks (right axis). Right: The latent representation values distribution grouped by event type.

Real-World Datasets

To extensively evaluate real-world performance, we consider a wide range of real-world datasets, briefly summarized below: (i) COVID: A dataset of patients admitted to the DUHS with positive COVID-19 testing, to predict death or use of a ventilator. (ii) InP (O'Brien et al. 2020): An in-patient data from DUHS, to predict the risk of death or ICR transfers. (iii) SEER (Ries et al. 2007): A public dataset studying cancer survival among adults curated by the U.S. Surveillance, Epidemiology, and End Results (SEER) Program, here we use a 10-year follow-up breast cancer subcohort. Summary statistics of these three real-world datasets are given in Table 3. Note that InP and SEER are survival datasets, specifically SEER includes censored subjects. We follow the data pre-processing steps in (Xiu et al. 2020). To create outcome labels, we set a cut-off time to define an event of interest the same as in the ablation study, and exclude subjects censored before the cut-off time (less than 0.2%). Datasets have been randomly split into training, validation, and testing datasets with ratio 6:2:2. See the SM for details on data pre-processing.

Table 2 compares VIE to its counterparts, where the numbers are averaged over the bootstrap samples. We see the proposed VIE yields the best performance in almost all cases, and the lead is more significant with low event rates. Note that the one-class classification based DeepSVDD performs poorly, which implies treating rare events as outliers are inappropriate. Re-weighting and resampling based methods (IW, Focal) are less stable compared to those simple baselines (LASSO, MLP). The theoretically optimal LDAM works well in general, second only to VIE in most settings. To further demonstrate the stability of our method, we visualize the bootstrapped evaluation scores for the COVID dataset in Figure 3, and defer the additional cross-validation results to the SM. We see that VIE leads consistently.

We also verify empirically that the estimated GPD shape parameters ξ_{GPD} are mostly positive (see the SM), indicating heavier than Gaussian tails as we have hypothesized. In Figure 2, we visualize one such latent dimension from the InP dataset, along with the associated risk learned by AMNN. In this example, the tail part is heavier than Gaussian and is associated with elevated risk. See our SM for examples where the extended tail contributes to prohibit the event.

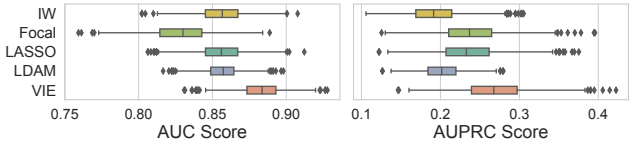

Figure 3: Bootstrapped AUC (left) and AUPRC (right) distributions for the COVID mortality data (2.6% event rate).

Conclusions

Motivated by the challenges of rare-event prediction in clinical settings, we presented Variational Inference with Extremals (VIE), a novel extreme representation learning-based variational solution to the problem. In this model we leveraged GPD to learn the extreme distributions with few samples and applied additive monotonic neural networks to disentangle the latent dimensions' effects on the outcome. VIE featured better generalization and interpretability, evidenced by strong performance on synthetic and real datasets.

Acknowledgements

The authors would like to thank the Duke Institute for Health Innovation (DIHI) for providing access to curated COVID-19 data and outcomes. This research was supported in part by NIH/NIDDK R01-DK123062, NIH/NIBIB R01-EB025020, NIH/NINDS 1R61NS120246-01, DARPA, DOE, ONR and NSF.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. In *ICLR*.

Aranda-Ordaz, F. J. 1981. On two families of transformations to additivity for binary response data. *Biometrika* 68(2): 357–363.

Bacchetti, P. 1989. Additive isotonic models. *Journal of the American Statistical Association* 84(405): 289–294.

Balkema, A. A.; and De Haan, L. 1974. Residual life time at great age. *The Annals of probability* 792–804.

Barlow, R. E.; Bartholomew, D. J.; Bremner, J. M.; and Brunk, H. D. 1972. Statistical inference under order restrictions: The theory and application of isotonic regression. Technical report, Wiley New York.

Bedoya, A. D.; Clement, M. E.; Phelan, M.; Steorts, R. C.; O'Brien, C.; and Goldstein, B. A. 2019. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical care medicine* 47(1): 49–55.

Bender, R.; Augustin, T.; and Blettner, M. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine* 24(11): 1713–1723.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8): 1798–1828.

Böhning, D.; Mylona, K.; and Kimber, A. 2015. Meta-analysis of clinical trials with rare events. *Biometrical Journal* 57(4): 633–648.

Botev, Z. I.; and Kroese, D. P. 2008. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability* 10(4): 471–505.

Breiman, L.; and Friedman, J. H. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80(391): 580–598.

Bryson, M. C. 1974. Heavy-tailed distributions: properties and tests. *Technometrics* 16(1): 61–68.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *ICML*.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*.

Chen, J.; Feng, J.; and Lu, W. 2018. A Wiener causality defined by relative entropy. In *ICONIP*.

Chen, J.; Xiu, Z.; Goldstein, B. A.; Henao, R.; Carin, L.; and Tao, C. 2020. Supercharging Imbalanced Data Learning With Causal Representation Transfer. *arXiv preprint arXiv:2011.12454* .

Chen, L.; Tao, C.; Zhang, R.; Henao, R.; and Carin, L. 2018. Variational inference and model selection with generalized evidence bounds. In *ICML*.

Coles, S.; Bawa, J.; Trenner, L.; and Dorazio, P. 2001. *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

Dai, B.; Dai, H.; He, N.; Liu, W.; Liu, Z.; Chen, J.; Xiao, L.; and Song, L. 2018. Coupled variational bayes via optimization embedding. In *NIPS*.

Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; and Bontempi, G. 2017. Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems* 29(8): 3784–3797.

Davis, P. J.; and Rabinowitz, P. 2007. *Methods of numerical integration*. Courier Corporation.

Davison, A. C.; and Smith, R. L. 1990. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)* 52(3): 393–425.

Drummond, C.; Holte, R. C.; et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *ICML Workshop*.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *IJCAI*.

Falk, M.; Hüsler, J.; and Reiss, R.-D. 2010. *Laws of small numbers: extremes and rare events*. Springer Science & Business Media.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80(1): 27–38.

Fithian, W.; and Hastie, T. 2014. Local case-control sampling: Efficient subsampling in imbalanced data sets. *The Annals of Statistics* 42(5): 1693.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2009. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* 3(4): 5.

Hahn, P.; and Jeruchim, M. 1987. Developments in the theory and application of importance sampling. *IEEE Transactions on Communications* 35(7): 706–714.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *ICIC*.

Hastie, T. J. 2017. Generalized additive models. In *Statistical models in S*, 249–307. Routledge.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263–1284.

He, J.; Spokoyny, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019. Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR*.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 5(1): 43–85.

Holte, R. C.; Acker, L.; Porter, B. W.; et al. 1989. Concept Learning and the Problem of Small Disjuncts. In *IJCAI*.

Huang, C.; Li, Y.; Change Loy, C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *CVPR*.

King, G.; and Zeng, L. 2001. Logistic regression in rare events data. *Political Analysis* 9(2): 137–163.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. In *NIPS*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *ICLR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Lu, D.; Guo, F.; and Li, F. 2020. Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods. *Accident Analysis & Prevention* 143: 105579.

Machado, J. T.; and Lopes, A. M. 2020. Rare and extreme events: the case of COVID-19 pandemic. *Nonlinear Dynamics* 1.

Mani, I.; and Zhang, I. 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. In *ICML Workshop*.

McNeil, A. J. 1997. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA* 27(1): 117–137.

Mescheder, L.; Nowozin, S.; and Geiger, A. 2017. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*.

Mitchell, G. F.; Hwang, S.-J.; Vasan, R. S.; Larson, M. G.; Pencina, M. J.; Hamburg, N. M.; Vita, J. A.; Levy, D.; and Benjamin, E. J. 2010. Arterial stiffness and cardiovascular events: the Framingham Heart Study. *Circulation* 121(4): 505.

Mitchell, T. M. 1999. Machine learning and data mining. *Communications of the ACM* 42(11): 30–36.

Munro, D.; Ersoy, O.; Bell, M.; and Sadowsky, J. 1996. Neural network learning of low-probability events. *IEEE Transactions on Aerospace and Electronic Systems* 32(3): 898–910.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

O'Brien, C.; Goldstein, B. A.; Shen, Y.; Phelan, M.; Lambert, C.; Bedoya, A. D.; and Steorts, R. C. 2020. Development, Implementation, and Evaluation of an In-Hospital Optimized Early Warning Score for Patient Deterioration. *MDM Policy & Practice* 5(1): 2381468319899663.

Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. In *NIPS*.

Pickands III, J.; et al. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* 3(1): 119–131.

Rezende, D. J.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *ICML*.

Ries, L. G.; Young, J.; Keel, G.; Eisner, M.; Lin, Y.; Horner, M.; et al. 2007. SEER survival monograph: cancer survival among adults: US SEER program, 1988-2001, patient and tumor characteristics. *National Cancer Institute, SEER Program, NIH Pub* (07-6215): 193–202.

Robert, C.; and Casella, G. 2013. *Monte Carlo statistical methods*. Springer Science & Business Media.

Rubinstein, R. Y.; and Kroese, D. P. 2013. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Rubinstein, R. Y.; and Kroese, D. P. 2016. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *ICML*.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2): 227–244.

Sill, J. 1998. Monotonic networks. In *NIPS*.

Sur, P.; and Candès, E. J. 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* 116(29): 14516–14525.

Tao, C.; Chen, L.; Dai, S.; Chen, J.; Bai, K.; Wang, D.; Feng, J.; Lu, W.; Bobashev, G.; and Carin, L. 2019. On Fenchel mini-max learning. In *NeurIPS*.

Tao, C.; Nichols, T. E.; Hua, X.; Ching, C. R.; Rolls, E. T.; Thompson, P. M.; Feng, J.; Initiative, A. D. N.; et al. 2017. Generalized reduced rank latent factor regression for high dimensional tensor fields, and neuroimaging-genetic applications. *NeuroImage* 144: 35–57.

Tax, D. M. J. 2002. *One-class classification: Concept learning in the absence of counter-examples*. Ph.D. thesis, Technische Universiteit Delft (The Netherlands).

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* 53(3): 1–34.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2017. Learning to model the tail. In *NIPS*.

Wehenkel, A.; and Louppe, G. 2019. Unconstrained monotonic neural networks. In *NeurIPS*.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10(2).

Xiu, Z.; Tao, C.; Goldstein, B. A.; and Henao, R. 2020. Variational learning of individual survival distributions. In *ACM CHIL*.

Yen, S.-J.; and Lee, Y.-S. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* 36(3): 5718–5727.

Zhang, L.; Zhang, C.; Quan, S.; Xiao, H.; Kuang, G.; and Liu, L. 2020. A Class Imbalance Loss for Imbalanced Object Recognition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 2778–2792.

Zhou, Z.-H.; and Liu, X.-Y. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* 18(1): 63–77.

Zong, W.; Huang, G.-B.; and Chen, Y. 2013. Weighted extreme learning machine for imbalance learning. *Neurocomputing* 101: 229–242.