Augment and Reduce: Stochastic Inference for Large Categorical Distributions

Francisco J. R. Ruiz [1 2] **Michalis K. Titsias** [3] **Adji B. Dieng** [2] **David M. Blei** [2]

Abstract

Categorical distributions are ubiquitous in machine learning, e.g., in classification, language models, and recommendation systems. However, when the number of possible outcomes is very large, using categorical distributions becomes computationally expensive, as the complexity scales linearly with the number of outcomes. To address this problem, we propose *augment and reduce* (A&R), a method to alleviate the computational complexity. A&R uses two ideas: latent variable augmentation and stochastic variational inference. It maximizes a lower bound on the marginal likelihood of the data. Unlike existing methods which are specific to softmax, A&R is more general and is amenable to other categorical models, such as multinomial probit. On several large-scale classification problems, we show that A&R provides a tighter bound on the marginal likelihood and has better predictive performance than existing approaches.

1. Introduction

Categorical distributions are fundamental to many areas of machine learning. Examples include classification (Gupta et al., 2014), language models (Bengio et al., 2006), recommendation systems (Marlin & Zemel, 2004), reinforcement learning (Sutton & Barto, 1998), and neural attention models (Bahdanau et al., 2015). They also play an important role in discrete choice models (McFadden, 1978).

A categorical is a die with K sides, a discrete random variable that takes on one of K unordered outcomes; a categorical distribution gives the probability of each possible outcome. Categorical variables are challenging to use when there are many possible outcomes. Such large categoricals appear in common applications such as image classification with many classes, recommendation systems with many items, and language models over large vocabularies. In this paper, we develop a new method for fitting and using large categorical distributions.

The most common way to form a categorical is through the softmax transformation, which maps a K-vector of reals to a distribution of K outcomes. Let ψ be a real-valued K-vector. The softmax transformation is

$$p(y = k \mid \psi) = \frac{\exp\{\psi_k\}}{\sum_{k'} \exp\{\psi_{k'}\}}. \tag{1}$$

Note the softmax is not the only way to map real vectors to categorical distributions; for example, the multinomial probit (Albert & Chib, 1993) is an alternative. Also note that in many applications, such as in multiclass classification, the parameter ψ_k is a function of per-sample features x. For example, a linear classifier forms a categorical over classes through a linear combination, $\psi_k = w_k^\top x$.

We usually fit a categorical with maximum likelihood estimation or any other closely related strategy. Given a dataset $y_{1:N}$ of categorical data—each y_n is one of K values—we aim to maximize the log likelihood,

$$\mathcal{L}_{\text{log likelihood}} = \sum_{n=1}^{N} \log p(y_n \mid \psi). \tag{2}$$

Fitting this objective requires evaluating both the log probability and its gradient.

Eqs. 1 and 2 reveal the challenge to using large categoricals. Evaluating the log probability and evaluating its gradient are both $\mathcal{O}(K)$ operations. But this is not OK: most algorithms for fitting categoricals—for example, stochastic gradient ascent—require repeated evaluations of both gradients and probabilities. When K is large, these algorithms are prohibitively expensive.

Here we develop a method for fitting large categorical distributions, including the softmax but also more generally. It is called augment and reduce (A&R). A&R rewrites the categorical distribution with an auxiliary variable ε,

$$p(y \mid \psi) = \int p(y, \varepsilon \mid \psi) d\varepsilon. \tag{3}$$

A&R then replaces the expensive log probability with a variational bound on the integral in Eq. 3. Using stochastic

[1]University of Cambridge. [2]Columbia University. [3]Athens University of Economics and Business.. Correspondence to: Francisco J. R. Ruiz <f.ruiz@eng.cam.ac.uk, f.ruiz@columbia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

variational methods (Hoffman et al., 2013), the cost to evaluate the bound (or its gradient) is far below $\mathcal{O}(K)$.

Because it relies on variational methods, A&R provides a lower bound on the marginal likelihood of the data. With this bound, we can embed A&R in a larger algorithm for fitting a categorical, e.g., a (stochastic) variational expectation maximization (VEM) algorithm (Beal, 2003). Though we focus on maximum likelihood, we can also use A&R in other algorithms that require $\log p(y \mid \psi)$ or its gradient, e.g., fully Bayesian approaches (Gelman et al., 2003) or the REINFORCE algorithm (Williams, 1992).

We study A&R on linear classification tasks with up to 10^4 classes. On simulated and real data, we find that it provides accurate estimates of the categorical probabilities and gives better performance than existing approaches.

Related work. There are many methods to reduce the cost of large categorical distributions, particularly under the softmax transformation. These include methods that approximate the exact computations (Gopal & Yang, 2013; Vijayanarasimhan et al., 2014), those that rely on sampling (Bengio & Sénécal, 2003; Mikolov et al., 2013; Devlin et al., 2014; Ji et al., 2016; Botev et al., 2017), those that use approximations and distributed computing (Grave et al., 2017), double-sum formulations (Raman et al., 2017; Fagan & Iyengar, 2018), and those that avail themselves of other techniques such as noise contrastive estimation (Smith & Jason, 2005; Gutmann & Hyvärinen, 2010) or random nearest neighbor search (Mussmann et al., 2017).

Other methods change the model. They might replace the softmax transformation with a hierarchical or stick-breaking model (Kurzynski, 1988; Morin & Bengio, 2005; Tsoumakas et al., 2008; Beygelzimer et al., 2009; Dembczyński et al., 2010; Khan et al., 2012). These approaches can be successful, but the structure of the hierarchy may influence the learned probabilities. Other methods replace the softmax with a scalable spherical family of losses (Vincent et al., 2015; de Brébisson & Vincent, 2016).

A&R is different from all of these techniques. Unlike many of them, it provides a lower bound on the log probability rather than an approximation. The bound is useful because it can naturally be embedded in algorithms like stochastic VEM. Further, the A&R methodology applies to transformations beyond the softmax. In this paper, we study large categoricals via softmax, multinomial probit, and multinomial logistic. A&R is the first scalable approach for the two latter models. It accelerates any transformation that can be recast as an additive noise model (e.g., Gumbel, 1954; Albert & Chib, 1993).

The approach that most closely relates to A&R is the one-vs-each (OVE) bound of Titsias (2016), which is a lower bound of the softmax. Like the other related methods, it is narrower than A&R in that it does not apply to transformations beyond the softmax. We also empirically compare A&R to OVE in Section 4. A&R provides a tighter lower bound and yields better predictive performance.

2. Augment and Reduce

We develop augment and reduce (A&R), a method for computing with large categorical random variables.

The utility perspective. A&R uses the additive noise model perspective on the categorical, which we refer to as the utility perspective. Define a *mean utility* ψ_k for each possible outcome $k \in \{1, \ldots, K\}$. To draw a variable y from a categorical, we draw a zero-mean noise term ε_k for each possible outcome and then choose the value that maximizes the *realized utility* $\psi_k + \varepsilon_k$. This corresponds to the following process,

$$\begin{aligned} \varepsilon_k &\sim \phi(\cdot), \quad k \in \{1, \ldots, K\}, \\ y &= \arg\max_k \left(\psi_k + \varepsilon_k\right). \end{aligned} \tag{4}$$

Note the errors ε_k are drawn fresh each time we draw a variable y. We assume that the errors are independent of each other, independent of the mean utility ψ_k, and identically distributed according to some distribution $\phi(\cdot)$.

Now consider the model where we marginalize the errors from Eq. 4. This results in a distribution $p(y \mid \psi)$, a categorical that transforms ψ to the simplex. Depending on the distribution of the errors, this induces different transformations. For example, a standard Gumbel distribution recovers the softmax transformation; a standard Gaussian recovers the multinomial probit transformation; a standard logistic recovers the multinomial logistic transformation.

Typically, the mean utility ψ_k is a function of observed features x, e.g., $\psi_k = x^\top w_k$ in linear models or $\psi_k = f_{w_k}(x)$ in non-linear settings. In both cases, w_k are model parameters, relating the features to mean utilities.

Let us focus momentarily on a linear classification problem under the softmax model. For each observation n, the mean utilities are $\psi_{nk} = x_n^\top w_k$ and the random errors ε_{nk} are Gumbel distributed. After marginalizing out the errors, the probability that observation n is in class k is given by Eq. 1, $p(y_n = k \mid x_n, w) \propto \exp\{x_n^\top w_k\}$. Fitting the classifier involves learning the weights w_k that parameterize ψ. For example, maximum likelihood uses gradient ascent to maximize $\sum_n \log p(y_n \mid x_n, w)$ with respect to w.

Large categoricals. When the number of outcomes K is large, the normalizing constant of the softmax is a computational burden; it is $\mathcal{O}(K)$. Consequently, it is burdensome to calculate useful quantities like $\log p(y_n \mid x_n, w)$ and its gradient $\nabla_w \log p(y_n \mid x_n, w)$. As an ultimate consequence, maximum likelihood estimation is slow—it needs to evalu-

ate the gradient for each n at each iteration.

Its difficulty scaling is not unique to the softmax. Similar issues arise for the multinomial probit and multinomial logistic. With these transformations as well, evaluating likelihoods and related quantities is $\mathcal{O}(K)$.

2.1. Augment and reduce

We introduce A&R to relieve this burden. A&R accelerates training in models with categorical distributions and a large number of outcomes.

Rather than operating directly on the marginal $p(y \mid \psi)$, A&R *augments* the model with one of the error terms and forms a joint $p(y, \varepsilon \mid \psi)$. (We drop the subscript n to avoid cluttered notation.) This augmented model has a desirable property: its log-joint is a sum over all the possible outcomes. A&R then *reduces*—it subsamples a subset of outcomes to construct estimates of the log-joint and its gradient. As a result, its complexity relates to the size of the subsample, not the total number of outcomes K.

The augmented model. Let $\phi(\varepsilon)$ be the distribution over the error terms, and $\Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \phi(\tau) d\tau$ the corresponding cumulative distribution function (CDF). The marginal probability of outcome k is the probability that its realized utility $(\psi_k + \varepsilon_k)$ is greater than all others,

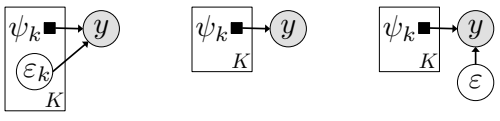
$$p(y = k \mid \psi) = \Pr\left(\psi_k + \varepsilon_k \geq \psi_{k'} + \varepsilon_{k'} \;\; \forall k' \neq k\right).$$

We write this probability as an integral over the kth error ε_k using the CDF of the other errors,

$$\begin{aligned} p(y = k \mid \psi) &= \int_{-\infty}^{+\infty} \phi(\varepsilon_k) \left(\prod_{k' \neq k} \int_{-\infty}^{\varepsilon_k + \psi_k - \psi_{k'}} \phi(\varepsilon_{k'}) d\varepsilon_{k'} \right) d\varepsilon_k \\ &= \int_{-\infty}^{+\infty} \phi(\varepsilon) \left(\prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'}) \right) d\varepsilon. \end{aligned} \tag{5}$$

(We renamed the dummy variable ε_k as ε to avoid clutter.) Eq. 5 is the same as found by Girolami & Rogers (2006) for the multinomial probit model, although we do not assume a Gaussian density $\phi(\varepsilon)$. Rather, we only assume that we can evaluate both $\phi(\varepsilon)$ and $\Phi(\varepsilon)$.

We derived Eq. 5 from the utility perspective, which encompasses many common models. We obtain the softmax by choosing a standard Gumbel distribution for $\phi(\varepsilon)$, in which case Eqs. 1 and 5 are equivalent. We obtain the multinomial probit by choosing a standard Gaussian distribution over the errors, and in this case the integral in Eq. 5 does not have a closed form. Similarly, we obtain the multinomial logistic by choosing a standard logistic distribution $\phi(\varepsilon)$. What is important is that regardless of the model, the cost to compute the marginal probability $p(y = k \mid \psi)$ is $\mathcal{O}(K)$.



(a) *(b)* *(c)*

Figure 1. (a) Illustration of the parameterization of a categorical model in terms of the utilities $\psi_k + \varepsilon_k$, where ψ_k is the mean utility and ε_k is an error term. The observed outcome is $y = \arg\max_k(\psi_k + \varepsilon_k)$. *(b)* In this model, the error terms have been marginalized out. This is the most common model for categorical distributions; it includes the softmax and multinomial probit. *(c)* The augmented model that we consider for A&R. All error terms have been integrated out, except one. In this model, the log-joint involves a summation over the possible outcomes k, enabling fast unbiased estimates of the log probability and its gradient.

We now augment the model with the auxiliary latent variable ε to form the joint distribution $p(y, \varepsilon \mid \psi)$,

$$p(y = k, \varepsilon \mid \psi) = \phi(\varepsilon) \prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'}). \tag{6}$$

This is a model that includes the kth error term from Eq. 4 but marginalizes out all the other errors. By construction, marginalizing ε from Eq. 6 recovers the original model $p(y \mid \psi)$ in Eq. 5. Figure 1 illustrates this idea.

Riihimäki et al. (2013) used Eq. 6 in the nested expectation propagation for Gaussian process classification. We use it to scale learning with categorical distributions.

The variational bound. The augmented model in Eq. 6 involves one latent variable ε. But our goal is to calculate the marginal $\log p(y \mid \psi)$ and its gradient. A&R derives a variational lower bound on $\log p(y \mid \psi)$ using the joint in Eq. 6. Define $q(\varepsilon)$ to be a variational distribution on the auxiliary variable. The bound is $\log p(y \mid \psi) \geq \mathcal{L}$, where

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\varepsilon)}\Big[\log p(y = k, \varepsilon \mid \psi) - \log q(\varepsilon)\Big] \\ &= \mathbb{E}_{q(\varepsilon)}\Big[\log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'}) - \log q(\varepsilon)\Big]. \end{aligned} \tag{7}$$

In Eq. 7, $\mathcal{L}$ is the evidence lower bound (ELBO); it is tight when $q(\varepsilon)$ is equal to the posterior of ε given y, $p(\varepsilon \mid y, \psi)$ (Jordan et al., 1999; Blei et al., 2017).

The ELBO contains a summation over the outcomes $k' \neq k$. A&R exploits this property to reduce complexity, as we describe below. Next we show how to use the bound in a variational expectation maximization (VEM) procedure and we describe the reduce step of A&R.

Variational expectation maximization. Consider again a linear classification task, where we have a dataset of features x_n and labels $y_n \in \{1, \ldots, K\}$ for $n = 1, \ldots, N$. The mean utility for each observation n is $\psi_{nk} = w_k^\top x_n$, and the goal is to learn the weights w_k by maximizing the log likelihood $\sum_n \log p(y_n \mid x_n, w)$.

A&R replaces each term in the data log likelihood with its bound using Eq. 7. The objective becomes $\sum_n \mathcal{L}^{(n)}$. Maximizing this objective requires an iterative process with two steps. In one step, A&R optimizes the objective with respect to w. In the other step, A&R optimizes each $\mathcal{L}^{(n)}$ with respect to the variational distribution. The resulting procedure takes the form of a VEM algorithm (Beal, 2003).

The VEM algorithm requires optimizing the ELBO with respect to w and the variational distributions.[1] This is challenging for two reasons. First, the expectations in Eq. 7 might not be tractable. Second, the cost to compute the gradients of Eq. 7 is still $\mathcal{O}(K)$.

Section 3 addresses these issues. To sidestep the intractable expectations, A&R forms unbiased Monte Carlo estimates of the gradient of the ELBO. To alleviate the computational complexity, A&R uses stochastic optimization, subsampling a set of outcomes k'.

Reduce by subsampling. The subsampling step in the VEM procedure is one of the key ideas behind A&R. Since Eq. 7 contains a summation over the outcomes $k' \neq k$, we can apply stochastic optimization techniques to obtain unbiased estimates of the ELBO and its gradient.

More specifically, consider the gradient of the ELBO in Eq. 7 with respect to w (the parameters of ψ). It is

$$\nabla_w \mathcal{L} = \sum_{k' \neq k} \mathbb{E}_{q(\varepsilon)} \big[\nabla_w \log \Phi(\varepsilon + \psi_k - \psi_{k'}) \big].$$

A&R estimates this by first randomly sampling a subset of outcomes $\mathcal{S} \subseteq \{1, \dots, K\} \setminus \{k\}$ of size $|\mathcal{S}|$. A&R then uses the outcomes in $\mathcal{S}$ to approximate the gradient,

$$\widetilde{\nabla}_w \mathcal{L} = \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \mathbb{E}_{q(\varepsilon)} \big[\nabla_w \log \Phi(\varepsilon + \psi_k - \psi_{k'}) \big].$$

This is an unbiased estimator[2] of the gradient $\nabla_w \mathcal{L}$. Crucially, A&R only needs to iterate over $|\mathcal{S}|$ outcomes to obtain it, reducing the complexity to $\mathcal{O}(|\mathcal{S}|)$.

The reduce step is also applicable to optimize the ELBO with respect to $q(\varepsilon)$. Section 3 gives further details about the stochastic VEM procedure in different settings.

3. Algorithm Description

Here we provide the details to run the variational expectation maximization (VEM) algorithm for the softmax model (Section 3.1) and for more general models including the multinomial probit and multinomial logistic (Section 3.2). These models only differ in the prior over the errors $\phi(\varepsilon)$.

Augment and reduce (A&R) is not limited to point-mass estimation of the parameters w. It is straightforward to extend the algorithm to perform posterior inference on w via stochastic variational inference, but for simplicity we describe maximum likelihood estimation.

3.1. Augment and Reduce for Softmax

In the softmax model, the distribution over the error terms is a standard Gumbel (Gumbel, 1954),

$$\phi_{\text{softmax}}(\varepsilon) = \exp\{-\varepsilon - e^{-\varepsilon}\}, \quad \Phi_{\text{softmax}}(\varepsilon) = \exp\{-e^{-\varepsilon}\}.$$

In this model, the optimal distribution $q^\star(\varepsilon)$, which achieves equality in the bound, has closed-form expression:

$$q^\star_{\text{softmax}}(\varepsilon) = \text{Gumbel}(\varepsilon\,;\, \log \eta^\star, 1),$$

with $\eta^\star = 1 + \sum_{k' \neq k} e^{\psi_{k'} - \psi_k}$. However, even though $q^\star_{\text{softmax}}(\varepsilon)$ has an analytic form, its parameter $\eta^\star$ is computationally expensive to obtain because it involves a summation over $K-1$ classes. Instead, we set

$$q_{\text{softmax}}(\varepsilon\,;\, \eta) = \text{Gumbel}(\varepsilon\,;\, \log \eta, 1).$$

Substituting this choice for $q_{\text{softmax}}(\varepsilon\,;\, \eta)$ into Eq. 7 gives the following evidence lower bound (ELBO):

$$\mathcal{L}_{\text{softmax}} = 1 - \log(\eta) - \frac{1}{\eta} \left(1 + \sum_{k' \neq k} e^{\psi_{k'} - \psi_k} \right). \quad (8)$$

Eq. 8 coincides with the log-concavity bound (Bouchard, 2007; Blei & Lafferty, 2007), although we have derived it from a completely different perspective. This derivation allows us to optimize η efficiently, as we describe next.

The Gumbel$(\varepsilon\,;\, \log \eta, 1)$ is an exponential family distribution whose natural parameter is η. This allows us to use natural gradients in the stochastic inference procedure. A&R iterates between a local step, in which we update η, and a global step, in which we update the parameters ψ.

In the local step (E step), we optimize η by taking a step in the direction of the noisy natural gradient, yielding $\eta^{\text{new}} = (1-\alpha)\eta^{\text{old}} + \alpha\widetilde{\eta}$. Here, $\widetilde{\eta}$ is an estimate of the optimal natural parameter, which we obtain using a random set of outcomes, i.e., $\widetilde{\eta} = 1 + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} e^{\psi_{k'} - \psi_k}$, where $\mathcal{S} \subseteq \{1, \dots, K\} \setminus \{k\}$. The parameter α is the step size; it must satisfy the Robbins-Monro conditions (Robbins & Monro, 1951; Hoffman et al., 2013).

In the global step (M step), we take a gradient step with respect to w (the parameters of ψ), holding η fixed. Similarly, we can estimate the gradient of Eq. 8 with complexity $\mathcal{O}(|\mathcal{S}|)$ by leveraging stochastic optimization.

[1] Note that maximizing the ELBO in Eq. 7 with respect to the distribution $q(\varepsilon)$ is equivalent to minimizing the Kullback-Leibler divergence from $q(\varepsilon)$ to the posterior $p(\varepsilon \mid y, \psi)$.

[2] This is not the only way to construct an unbiased estimator. Alternatively, we can draw the outcomes k' using importance sampling, taking into account the frequency of each class. We leave this for future work.

Algorithm 1 Softmax A&R for classification

Input: data (x_n, y_n), minibatch sizes $|\mathcal{B}|$ and $|\mathcal{S}|$
Output: weights $w = \{w_k\}_{k=1}^K$
Initialize all weights and natural parameters
for iteration $t = 1, 2, \ldots,$ **do**
  `# Sample minibatches:`
  Sample a minibatch of data, $\mathcal{B} \subseteq \{1, \ldots, N\}$
  for $n \in \mathcal{B}$ **do**
    Sample a set of labels, $\mathcal{S}_n \subseteq \{1, \ldots, K\} \setminus \{y_n\}$
  end for
  `# Local step (E step):`
  for $n \in \mathcal{B}$ **do**
    Compute $\widetilde{\eta}_n = 1 + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}_n} e^{\psi_{nk'} - \psi_{ny_n}}$
    Update natural param., $\eta_n \leftarrow (1 - \alpha^{(t)})\eta_n + \alpha^{(t)} \widetilde{\eta}_n$
  end for
  `# Global step (M step):`
  Set $g = -\frac{N}{|\mathcal{B}|} \frac{K-1}{|\mathcal{S}|} \sum_{n \in \mathcal{B}} \frac{1}{\eta_n} \sum_{k' \in \mathcal{S}_n} \nabla_w e^{\psi_{nk'} - \psi_{ny_n}}$
  Gradient step on the weights, $w \leftarrow w + \rho^{(t)} g$
end for

Algorithm 1 summarizes the procedure for a classification task. In this example, the dataset consists of N datapoints (x_n, y_n), where x_n is a feature vector and $y_n \in \{1, \ldots, K\}$ is the class label. Each observation is associated with its parameters ψ_{nk}; e.g., $\psi_{nk} = x_n^\top w_k$. We posit a softmax likelihood, and we wish to infer the weights via maximum likelihood using A&R. Thus, the objective function is $\sum_n \mathcal{L}_{\text{softmax}}^{(n)}$. (It is straightforward to obtain the maximum a posteriori solution by adding a regularizer.) At each iteration, we process a random subset of observations as well as a random subset of classes for each one.

Finally, note that we can perform posterior inference on the parameters w (instead of maximum likelihood) using A&R. One way is to consider a variational distribution $q(w)$ and take gradient steps with respect to the variational parameters of $q(w)$ in the global step, using the reparameterization trick (Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014; Kingma & Welling, 2014) to approximate that gradient. In the local step, we only need to evaluate the moment generating function, estimating the optimal natural parameter as $\widetilde{\eta} = 1 + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \mathbb{E}_{q(w)} \left[e^{\psi_{k'} - \psi_k} \right]$.

3.2. Augment and Reduce for Other Models

For most models, the expectations of the ELBO in Eq. 7 are intractable, and there is no closed-form solution for the optimal variational distribution $q^\star(\varepsilon)$. Fortunately, we can apply A&R, using the reparameterization trick to build Monte Carlo estimates of the gradient of the ELBO with respect to the variational parameters (Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014; Kingma & Welling, 2014).

More in detail, consider the variational distribution $q(\varepsilon\,;\,\nu)$, parameterized by some variational parameters ν. We assume that this distribution is reparameterizable, i.e., we can sample from $q(\varepsilon\,;\,\nu)$ by first sampling an auxiliary variable $u \sim q^{(\text{rep})}(u)$ and then setting $\varepsilon = T(u\,;\,\nu)$.

Algorithm 2 General A&R for classification

Input: data (x_n, y_n), minibatch sizes $|\mathcal{B}|$ and $|\mathcal{S}|$
Output: weights $w = \{w_k\}_{k=1}^K$
Initialize all weights and local variational parameters
for iteration $t = 1, 2, \ldots,$ **do**
  `# Sample minibatches:`
  Sample a minibatch of data, $\mathcal{B} \subseteq \{1, \ldots, N\}$
  for $n \in \mathcal{B}$ **do**
    Sample a set of labels, $\mathcal{S}_n \subseteq \{1, \ldots, K\} \setminus \{y_n\}$
  end for
  `# Local step (E step):`
  for $n \in \mathcal{B}$ **do**
    Sample auxiliary variable $u_n \sim q^{(\text{rep})}(u_n)$
    Transform auxiliary variable, $\varepsilon_n = T(u_n\,;\,\nu_n)$
    Estimate the gradient $\widetilde{\nabla}_{\nu_n} \mathcal{L}^{(n)}$ (Eq. 9)
    Update variational param., $\nu_n \leftarrow \nu_n + \alpha^{(t)} \widetilde{\nabla}_{\nu_n} \mathcal{L}^{(n)}$
  end for
  `# Global step (M step):`
  Sample $\varepsilon_n \sim q(\varepsilon_n\,;\,\nu_n)$ for all $n \in \mathcal{B}$
  Set $g = \frac{N}{|\mathcal{B}|} \frac{K-1}{|\mathcal{S}|} \sum_{n \in \mathcal{B}} \sum_{k' \in \mathcal{S}_n} \nabla_w \log \Phi(\varepsilon_n + \psi_{ny_n} - \psi_{nk'})$
  Gradient step on the weights, $w \leftarrow w + \rho^{(t)} g$
end for

In the local step, we fit $q(\varepsilon\,;\,\nu)$ by taking a gradient step of the ELBO with respect to the variational parameters ν. Since the expectations in Eq. 7 are not tractable, we obtain Monte Carlo estimates by sampling ε from the variational distribution. To sample ε, we sample $u \sim q^{(\text{rep})}(u)$ and set $\varepsilon = T(u\,;\,\nu)$. To alleviate the computational complexity, we apply the reduce step, sampling a random subset $\mathcal{S} \subseteq \{1, \ldots, K\} \setminus \{k\}$ of outcomes. We thus form a one-sample gradient estimator as

$$\widetilde{\nabla}_\nu \mathcal{L} = \nabla_\varepsilon \log \widetilde{p}(y, \varepsilon \,|\, \psi) \nabla_\nu T(u\,;\,\nu) + \nabla_\nu \mathbb{H}[q(\varepsilon\,;\,\nu)], \tag{9}$$

where $\mathbb{H}[q(\varepsilon\,;\,\nu)]$ is the entropy of the variational distribution,[3] and $\log \widetilde{p}(y, \varepsilon \,|\, \psi)$ is a log joint estimate,

$$\log \widetilde{p}(y, \varepsilon \,|\, \psi) = \log \phi(\varepsilon) + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \log \Phi(\varepsilon + \psi_k - \psi_{k'}).$$

In the global step, we estimate the gradient of the ELBO with respect to w. Following a similar approach, we obtain an unbiased one-sample gradient estimator as $\widetilde{\nabla}_w \mathcal{L} = \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \nabla_w \log \Phi(\varepsilon + \psi_k - \psi_{k'})$.

Algorithm 2 summarizes the procedure to efficiently run

[3] We can estimate the gradient of the entropy when it is not available analytically. Even when it is, the Monte Carlo estimator may have lower variance (Roeder et al., 2017).

maximum likelihood on a classification problem. We subsample observations and classes at each iteration.

Finally, note that we can perform posterior inference on the parameters w by positing a variational distribution $q(w)$ and taking gradient steps with respect to the variational parameters of $q(w)$ in the global step. In this case, the reparameterization trick is needed in both the local and global step to obtain Monte Carlo estimates of the gradient.

We now particularize A&R for the multinomial probit and multinomial logistic models.

A&R for multinomial probit. Consider a standard Gaussian distribution over the error terms,

$$\phi_{\text{probit}}(\varepsilon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\varepsilon^2}, \quad \Phi_{\text{probit}}(\varepsilon) = \int_{-\infty}^{\varepsilon} \phi_{\text{probit}}(\tau) d\tau.$$

A&R chooses a Gaussian variational distribution $q_{\text{probit}}(\varepsilon\,;\,\nu) = \mathcal{N}(\varepsilon\,;\,\mu, \sigma^2)$ and fits the variational parameters $\nu = [\mu,\ \sigma]^\top$. The Gaussian is reparameterizable in terms of a standard Gaussian, i.e., $q^{(\text{rep})}_{\text{probit}}(u) = \mathcal{N}(u\,;\,0, 1)$. The transformation is $\varepsilon = T(u\,;\,\nu) = \mu + \sigma u$. Thus, the gradients in Eq. 9 are $\nabla_\nu T(u\,;\,\nu) = [1,\ u]^\top$ and $\nabla_\nu \mathbb{H}[q_{\text{probit}}(\varepsilon\,;\,\nu)] = [0,\ 1/\sigma]^\top$.

A&R for multinomial logistic. Consider now a standard logistic distribution over the errors,

$$\phi_{\text{logistic}}(\varepsilon) = \sigma(\varepsilon)\sigma(-\varepsilon), \qquad \Phi_{\text{logistic}}(\varepsilon) = \sigma(\varepsilon),$$

where $\sigma(\varepsilon) = \frac{1}{1+e^{-\varepsilon}}$ is the sigmoid function. (The logistic distribution has heavier tails than the Gaussian.) Under this model, the ELBO in Eq. 7 takes the form

$$\mathcal{L}_{\text{logistic}} = \mathbb{E}_{q(\varepsilon)}\Big[\log \frac{\sigma(\varepsilon)\sigma(-\varepsilon)}{q(\varepsilon)} + \sum_{k'\neq k} \log \sigma(\varepsilon + \psi_k - \psi_{k'})\Big].$$

Note the close resemblance between this expression and the one-vs-each (OVE) bound of Titsias (2016),

$$\mathcal{L}_{\text{OVE}} = \sum_{k'\neq k} \log \sigma(\psi_k - \psi_{k'}). \tag{10}$$

However, while the former is a bound on the multinomial logistic model, the OVE is a bound on the softmax.

A&R sets $q_{\text{logistic}}(\varepsilon\,;\,\nu) = \frac{1}{\beta}\sigma\Big(\frac{\varepsilon-\mu}{\beta}\Big)\sigma\Big(-\frac{\varepsilon-\mu}{\beta}\Big)$, a logistic distribution. The variational parameters are $\nu = [\mu,\ \beta]^\top$. The logistic distribution is reparameterizable, with $q^{(\text{rep})}_{\text{logistic}}(u) = \sigma(u)\sigma(-u)$ and transformation $\varepsilon = T(u\,;\,\nu) = \mu + \beta u$. The gradient of the entropy in Eq. 9 is $\nabla_\nu \mathbb{H}[q_{\text{logistic}}(\varepsilon\,;\,\nu)] = [0,\ 1/\beta]^\top$.

4. Experiments

We showcase augment and reduce (A&R) on a linear classification task. Our goal is to assess the predictive performance of A&R in this classification task, to assess the quality of the marginal bound of the data, and to compare its complexity[4] with existing approaches.

We run A&R for three different models of categorical distributions (softmax, multinomial probit, and multinomial logistic).[5] For the softmax model, we compare A&R against the one-vs-each (OVE) bound (Titsias, 2016). Just like A&R, OVE is a rigorous lower bound on the marginal likelihood. It can also run on a single machine,[6] and it has been shown to outperform other approaches.

For softmax, A&R runs nearly as fast as OVE but has better predictive performance and provides a tighter bound on the marginal likelihood than OVE. On two small datasets, the A&R bound closely reaches the marginal likelihood of exact softmax maximum likelihood estimation.

We now describe the experimental settings. In Section 4.1, we analyze synthetic data and $K = 10^4$ classes. In Section 4.2, we analyze five real datasets.

Experimental setup. We consider linear classification, where the mean utilities are $\psi_{nk} = w_k^\top x_n + w_k^{(0)}$. We fit the model parameters (weights and biases) via maximum likelihood estimation, using stochastic gradient ascent. We initialize the weights and biases randomly, drawing from a Gaussian distribution with zero mean and standard deviation 0.1 (0.001 for the biases). For each experiment, we use the same initialization across all methods.

Algorithms 1 and 2 require setting a step size schedule for $\rho^{(t)}$. We use the adaptive step size sequence proposed by Kucukelbir et al. (2017), which combines RMSPROP (Tieleman & Hinton, 2012) and Adagrad (Duchi et al., 2011). We set the step size using the default parameters, i.e.,

$$\rho^{(t)} = \rho_0 \times t^{-1/2+10^{-16}} \times \Big(1 + \sqrt{s^{(t)}}\Big)^{-1},$$
$$s^{(t)} = 0.1(g^{(t)})^2 + 0.9 s^{(t-1)}.$$

We set $\rho_0 = 0.02$ and we additionally decrease ρ_0 by a factor of 0.9 every 2000 iterations. We use the same step size sequence for OVE.

We set the step size $\alpha^{(t)}$ in Algorithm 1 as $\alpha^{(t)} = (1+t)^{-0.9}$, the default values suggested by Hoffman et al. (2013). For the step size $\alpha^{(t)}$ in Algorithm 2, we set $\alpha^{(t)} = 0.01(1+t)^{-0.9}$. For the multinomial logit and multinomial probit A&R, we parameterize the variational distributions in terms of their means μ and their unconstrained scale parameter γ, such that the scale parameter is $\log(1+\exp(\gamma))$.

[4] We focus on runtime cost. A&R requires $\mathcal{O}(N)$ memory storage capacity due to the local variational parameters.

[5] Code for A&R is available at https://github.com/franrruiz/augment-reduce.

[6] A&R is amenable to an embarrassingly parallel algorithm, but we focus on single-core procedures.

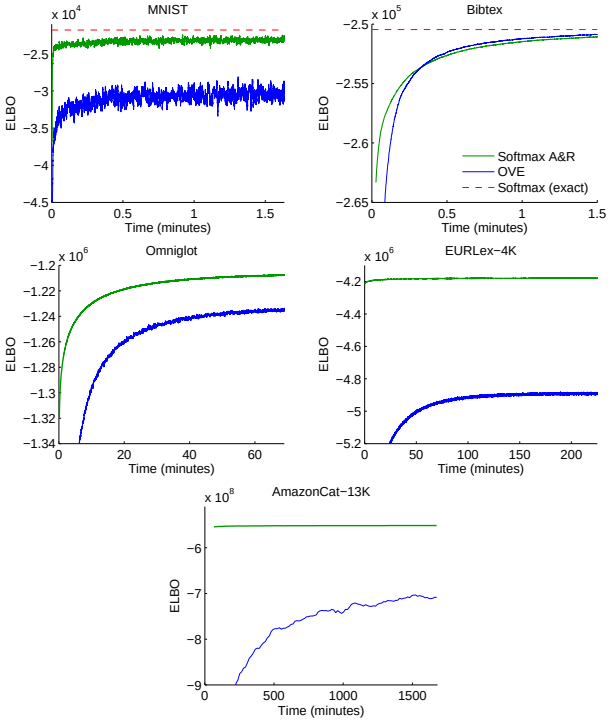


Figure 2. Evolution of the ELBO as a function of wall-clock time. The softmax A&R method provides a tighter bound than OVE (Titsias, 2016) for almost all the considered datasets.

4.1. Synthetic Dataset

We mimic the toy experiment of Titsias (2016) to assess how well A&R estimates the categorical probabilities. We generate a dataset with 10^4 classes and $N = 3 \times 10^5$ observations, each assigned label k with probability $p_k \propto \widetilde{p}_k^2$, where each $\widetilde{p}_k$ is randomly generated from a uniform distribution in $[0, 1]$. After generating the data, we have $K = 9{,}035$ effective classes (thus we use this value for K). In this simple setting, there are no observed covariates x_n.

We estimate the probabilities p_k via maximum likelihood on the biases $w_k^{(0)}$. We posit a softmax model, and we apply both the variational expectation maximization (VEM) in Section 3.1 and the OVE bound. For both approaches, we choose a minibatch size of $|\mathcal{B}| = 500$ observations and $|\mathcal{S}| = 100$ classes, and we run 5×10^5 iterations.

We run each approach on one CPU core. On average, the wall-clock time per epoch (one epoch takes $N/|\mathcal{B}| = 600$ iterations) is 0.196 minutes for softmax A&R and 0.189 minutes for OVE. A&R is slightly slower because of the local step that OVE does not require; however, the bound on the marginal log likelihood is tighter (by orders of magnitude) for A&R than for OVE (-2.62×10^6 and -1.40×10^9, respectively). The estimated probabilities are similar for both methods: the average absolute error is 3.00×10^{-6} for A&R and 3.65×10^{-6} for OVE; the difference is not statistically significant.

4.2. Real Datasets

We now turn to real datasets. We consider MNIST and Bibtex (Katakis et al., 2008; Prabhu & Varma, 2014), where we can compare against the exact softmax. We also analyze Omniglot (Lake et al., 2015), EURLex-4K (Mencia & Furnkranz, 2008; Bhatia et al., 2015), and AmazonCat-13K (McAuley & Leskovec, 2013).[7] Table 1 gives information about the structure of these datasets.

We run each method for a fixed number of iterations. We set the minibatch sizes $|\mathcal{B}|$ and $|\mathcal{S}|$ beforehand. The specific values for each dataset are also in Table 1.

Data preprocessing. For MNIST, we divide the pixel values by 255 so that the maximum value is one. For Omniglot, following other works in the literature (e.g., Burda et al., 2016), we resize the images to 28×28 pixels. For EURLex-4K and AmazonCat-13K, we normalize the covariates dividing by their maximum value.

Bibtex, EURLex-4K, and AmazonCat-13K are multi-class datasets, i.e., each observation may be assigned more than one label. Following Titsias (2016), we keep only the first non-zero label for each data point. See Table 1 for the resulting number of classes in each case.

Evaluation. For the softmax, we compare A&R against the OVE bound.[8] We also compare against the exact softmax on MNIST and Bibtex, where the number of classes is small. For the multinomial probit and multinomial logistic models, we also report the predictive performance of A&R.

We evaluate performance with test log likelihood and accuracy. The accuracy is the fraction of correctly classified instances, assuming that we assign the most likely label (i.e., the one with the highest mean utility). To compute the test log likelihood, we use Eq. 1 for the softmax and Eq. 5 for the multinomial probit and multinomial logistic models. We approximate the integral in Eq. 5 with 1,000 samples using importance sampling (we use a Gaussian distribution with mean 5 and standard deviation 5 as a proposal).

Results. Table 2 shows the wall-clock time per epoch for each method and dataset. In general, softmax A&R is almost as fast as OVE because the extra local step can be performed efficiently without additional expensive operations. It requires to evaluate exponential functions that can be reused in the global step. Multinomial probit A&R and multinomial

[7] MNIST is available at http://yann.lecun.com/exdb/mnist. Omniglot can be found at https://github.com/brendenlake/omniglot. Bibtex, EURLex-4K, and AmazonCat-13K are available at http://manikvarma.org/downloads/XC/XMLRepository.html.

[8] We also implemented the approach of Botev et al. (2017), but we do not report the results because it did not outperform OVE in terms of test log-likelihood on four out of the five considered datasets. On the fifth dataset, softmax A&R was still superior.

Table 1. Statistics and experimental settings of the considered datasets. N_{train} and N_{test} are the number of training and test data points. The number of classes is the resulting value after the preprocessing step (see text). The minibatch sizes correspond to $|\mathcal{B}|$ and $|\mathcal{S}|$, respectively.

dataset	N_{train}	N_{test}	covariates	classes	minibatch (obs.)	minibatch (classes)	iterations
MNIST	$60,000$	$10,000$	784	10	500	1	$35,000$
Bibtex	$4,880$	$2,413$	$1,836$	148	488	20	$5,000$
Omniglot	$25,968$	$6,492$	784	$1,623$	541	50	$45,000$
EURLex-4K	$15,539$	$3,809$	$5,000$	896	379	50	$100,000$
AmazonCat-13K	$1,186,239$	$306,782$	$203,882$	$2,919$	$1,987$	60	$5,970$

Table 2. Average time per epoch for each method and dataset. Softmax A&R (Section 3.1) is almost as fast as OVE. The A&R approaches in Section 3.2 take longer because they require some additional computations, but they are still competitive.

		A&R [this paper]		
dataset	OVE (Titsias, 2016)	softmax	multi. probit	multi. logistic
MNIST	0.336 s	0.337 s	0.431 s	0.511 s
Bibtex	0.181 s	0.188 s	0.244 s	0.246 s
Omniglot	4.47 s	4.65 s	5.63 s	5.57 s
EURLex-4K	5.54 s	5.65 s	6.46 s	6.23 s
AmazonCat-13K	2.80 h	2.80 h	2.82 h	2.91 h

Table 3. Test log likelihood and accuracy for each method and dataset. The table on the left compares the approaches based on the softmax. Softmax A&R outperforms OVE in four out of the five datasets. The two tables on the right show the performance of other models (multinomial probit and multinomial logistic), for which A&R is also competitive.

			softmax model				multi. probit		multi. logistic	
	exact		OVE (Titsias, 2016)		A&R [this paper]		A&R [this paper]		A&R [this paper]	
dataset	log lik	acc	log lik	acc	log lik	acc	log lik	acc	log lik	acc
MNIST	-0.261	0.927	-0.276	0.919	$\mathbf{-0.271}$	$\mathbf{0.924}$	-0.302	0.918	-0.287	0.917
Bibtex	-3.188	0.361	-3.300	0.352	$\mathbf{-3.036}$	$\mathbf{0.361}$	-4.184	0.346	-3.151	0.353
Omniglot	–	–	-5.667	0.179	$\mathbf{-5.171}$	$\mathbf{0.201}$	-7.350	0.178	-5.395	0.184
EURLex-4K	–	–	$\mathbf{-4.241}$	$\mathbf{0.247}$	-4.593	0.207	-4.193	0.263	-4.299	0.226
AmazonCat-13K	–	–	-3.880	0.388	$\mathbf{-3.795}$	$\mathbf{0.420}$	-3.593	0.411	-4.081	0.350

logistic A&R are slightly slower because of the local step, but they are still competitive.

For the five datasets, Figure 2 shows the evolution of the evidence lower bound (ELBO) as a function of wall-clock time for the softmax A&R (Eq. 8), compared to the OVE (Eq. 10). For easier visualization, we plot a smoothed version of the bounds after applying a moving average window of size 100. (For AmazonCat-13K, we only compute the ELBO every 50 iterations and we use a window of size 5.) Softmax A&R provides a significantly tighter bound for most datasets (except for Bibtex, where the ELBO of A&R is close to the OVE bound). For MNIST and Bibtex, we also plot the marginal likelihood obtained after running maximum likelihood estimation on the exact softmax model. The ELBO of A&R nearly achieves this value.

Finally, Table 3 shows the predictive performance for all methods across all datasets. We report test log likelihood and accuracy. Softmax A&R outperforms OVE in both metrics on all but one dataset (except EURLex-4K). Although our goal is not to compare performance across different models, for completeness Table 3 also shows the predictive performance of multinomial probit A&R and multinomial logistic A&R. In general, softmax A&R provides the highest test log likelihood, but multinomial probit A&R outperforms all other methods in EURLex-4K and AmazonCat-13K. Additionally, multinomial logistic A&R presents better predictive performance than OVE on Omniglot and Bibtex.

5. Conclusion

We have introduced augment and reduce (A&R), a scalable method to fit models involving categorical distributions. A&R is general and applicable to many models, including the softmax and the multinomial probit. On classification tasks, we found that A&R outperforms state-of-the art algorithms with little extra computational cost.

Acknowledgements

This work was supported by ONR N00014-15-1-2209, ONR 133691-5102004, NIH 5100481-5500001084, NSF CCF-1740833, the Alfred P. Sloan Foundation, the John Simon Guggenheim Foundation, Facebook, Amazon, and IBM. Francisco J. R. Ruiz is supported by the EU Horizon 2020 programme (Marie Skłodowska-Curie Individual Fellowship, grant agreement 706760). We also thank Victor Elvira and Pablo Moreno for their comments and help.

References

Albert, J. H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Beal, M. J. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

Bengio, Y. and Sénécal, J.-S. Quick training of probabilistic neural nets by importance sampling. In *Artificial Intelligence and Statistics*, 2003.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. Neural probabilistic language models. In *Innovations in Machine Learning*. Springer, 2006.

Beygelzimer, A., Langford, J., Lifshits, Y., Sorkin, G. B., and Strehl, L. Conditional probability tree estimation analysis and algorithms. In *Uncertainty in Artificial Intelligence*, 2009.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*, 2015.

Blei, D., Kucukelbir, A., and McAuliffe, J. Variational inference: A review for statisticians. *Journal of American Statistical Association*, 2017.

Blei, D. M. and Lafferty, J. D. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

Botev, A., Zheng, B., and Barber, D. Complementary sum sampling for likelihood approximation in large scale classification. In *Artificial Intelligence and Statistics*, 2017.

Bouchard, G. Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *Advances in Neural Information Processing Systems, Workshop on Approximate Inference in Hybrid Models*, 2007.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

de Brébisson, A. and Vincent, P. An exploration of softmax alternatives belonging to the spherical loss family. In *International Conference on Learning Representations*, 2016.

Dembczyński, K., Cheng, W., and Hüllermeier, E. Bayes optimal multilabel classification via probabilistic classifier chains. In *International Conference on Machine Learning*, 2010.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. Fast and robust neural network joint models for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, Jul 2011.

Fagan, F. and Iyengar, G. Unbiased scalable softmax optimization. In *arXiv:1803.08577*, 2018.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 2003.

Girolami, M. and Rogers, S. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

Gopal, S. and Yang, Y. Distributed training of large-scale logistic models. In *International Conference on Machine Learning*, 2013.

Grave, E., Joulin, A., Cissé, M., Grangier, D., and Jégrou, H. Efficient softmax approximation for GPUs. In *arXiv:1609.04309*, 2017.

Gumbel, E. J. *Statistical theory of extreme values and some practical applications: A series of lectures*. U. S. Govt. Print. Office, 1954.

Gupta, M. R., Bengio, S., and Jason, W. Training highly multiclass classifiers. *Journal of Machine Learning Research*, 15(1):1461–1492, 2014.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Artificial Intelligence and Statistics*, 2010.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, May 2013.

Ji, S., Vishwanathan, S. V. N., Satish, N., Anderson, M. J., and Dubey, P. Blackout: Speeding up recurrent neural network language models with very large vocabularies. In *International Conference on Learning Representations*, 2016.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical

models. *Machine Learning*, 37(2):183–233, November 1999.

Katakis, I., Tsoumakas, G., and Vlahavas, I. Multilabel text classification for automated tag suggestion. In *ECML/PKDD Discovery Challenge*, 2008.

Khan, M. E., Mohamed, S., Marlin, B. M., and Murphy, K. P. A stick-breaking likelihood for categorical data analysis with latent Gaussian models. In *Artificial Intelligence and Statistics*, 2012.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14): 1–45, 2017.

Kurzynski, M. On the multistage Bayes classifier. *Pattern Recognition*, 21(4):355–365, 1988.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Marlin, B. M. and Zemel, R. S. The multiple multiplicative factor model for collaborative filtering. In *International Conference on Machine Learning*, 2004.

McAuley, J. and Leskovec, J. Hidden factors and hidden topics: understanding rating dimensions with review text. In *ACM Conference on Recommender Systems*, 2013.

McFadden, D. Modeling the choice of residential location. In *Spatial Interaction Theory and Residential Location*, 1978.

Mencia, E. L. and Furnkranz, J. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *ECML/PKDD*, 2008.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

Morin, F. and Bengio, Y. Hierarchical probabilistic neural network language model. In *Artificial Intelligence and Statistics*, 2005.

Mussmann, S., Levy, D., and Ermon, S. Fast amortized inference and learning in log-linear models with randomly perturbed nearest neighbor search. In *Uncertainty in Artificial Intelligence*, 2017.

Prabhu, Y. and Varma, M. FastXML: Fast, accurate and stable tree-classifier for eXtreme multi-label learning. In *KDD*, 2014.

Raman, P., Matsushima, S., Zhang, X., Yun, H., and Vishwanathan, S. V. N. DS-MLR: exploiting double separability for scaling up distributed multinomial logistic regression. *arXiv:1604.04706v2*, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Riihimäki, J., Jylänki, P., and Vehtari, A. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Roeder, G., Wu, Y., and Duvenaud, D. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, 2017.

Smith, N. A. and Jason, E. Contrastive estimation: Training log-linear models on unlabeled data. In *Association for Computational Linguistics*, 2005.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Tieleman, T. and Hinton, G. Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 2012.

Titsias, M. K. One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*, 2016.

Titsias, M. K. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

Tsoumakas, G., Katakis, I., and Vlahavas, I. Effective and efficient multilabel classification in domains with large number of labels. In *ECML/PKDD Workshop on Mining Multidimensional Data*, 2008.

Vijayanarasimhan, S., Shlens, J., Monga, R., and Yagnik, J. Deep networks with large output spaces. In *arXiv:1412.7479*, 2014.

Vincent, P., de Brébisson, A., and Bouthillier, X. Efficient exact gradient update for training deep networks with very large sparse targets. In *Advances in Neural Information Processing Systems*, 2015.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.