Asymptotic Simulated Annealing for Variational Inference

San Gultekin
Columbia University
New York, NY, US
sg3108@columbia.edu

Aonan Zhang
Columbia University
New York, NY, US
az2385@columbia.edu

John Paisley
Columbia University
New York, NY, US
jpaisley@columbia.edu

***Abstract*—Variational inference (VI) is an effective deterministic method for approximate posterior inference, which arises in many practical applications. However, it typically suffers from non-convexity issues. This paper proposes a novel optimization tool called asymptotically-annealed variational inference (AVI), for better local optimal convergence of VI by using ideas from small-variance asymptotics to efficiently search for better solutions. The algorithm entails a simple modification to the basic VI algorithm, has little additional computational cost and is very simple. Furthermore, our algorithm can be viewed as an asymptotic limit of simulated annealing, connecting it to a recent literature in machine learning on deterministic versions of stochastic algorithms. Experiments show better convergence performance than VI and other annealing methods for models such as LDA and the HMM, as well as on stochastic variational inference problems for big data.**

***Index Terms*—Variational inference, simulated annealing, small variance asymptotics, big data.**

I. Introduction

The rapid growth of data is one of the big challenges faced by the contemporary machine learning. Scalable inference makes many applications in fields such as genetics, finance, and recommender systems possible. To avoid overfitting, Bayesian methods for machine learning can produce better models through a probabilistic framework that treats model parameters as random variables [1]. This advantage, however, is often faced by an intractable posterior computation.

There is a large literature on approximating posterior distributions of a model. The traditional method is Markov chain Monte Carlo (MCMC) which obtains samples from an approximation to the true posterior distribution by simulating a Markov chain [2]. While strong convergence results are available for such algorithms, scalability is often an issue. As an alternative, variational inference treats posterior inference as an optimization problem over distribution parameters [1]. Variational methods have been successfully applied in many settings including communication systems; for example multiple-access channels, multi-user detection, and OFDM systems [3]–[6]. Importantly, this approach results in much faster algorithms, which can be scaled to very large datasets using stochastic optimization [7]. However, as shown by [8], such algorithms are usually highly non-convex even for a small number of parameters. Modern applications are typically high-dimensional and contain many parameters, which makes local optima a major obstacle in inference.

Escaping bad local optima for non-convex optimization problems is a long-standing problem. Popular methods for addressing this issue are convex relaxations [8], swarm intelligence techniques such as particle swarm optimization [9], and simulated annealing [10]. However, not all of these techniques are equally effective for variational inference, considering scalability. In this context, deterministic annealing for variational inference has been particularly useful [11], [12]. The basic idea here is that, by scaling up the entropy term in the objective function, some local optima can be smoothed out for better convergence. The recently proposed variational tempering [13] extends this idea to stochastic variational inference (SVI) [7].

While having clear practical benefits, deterministic annealing comes with some drawbacks. First, this algorithm is unaware of the variational landscape in the sense that at each step the parameters are updated without relating the previous and current parameter values. Consequently, the choice of cooling schedule (i.e. the factor we use to dampen the objective function) becomes crucial and oftentimes the chosen cooling schedule can be inappropriate. Automatic temperature selection is proposed by [13], but this significantly increases the computational cost. On the other hand, the quantum annealing framework [14] uses different states that are linked through the variational E-step, also with increased computation. Another common drawback of these approaches is that they all require a re-derivation of the variational inference algorithm for each new model, which can hinder their automation and widespread use.

For these reasons, it is useful to develop an optimization technique which can be integrated into the variational inference procedure with minimal modification and negligible cost. One possible candidate is simulated annealing, but like MCMC this approach has drawbacks as it incurs significant computational cost [2]. In this paper we propose an asymptotic version of simulated annealing for variational inference (AVI) which is guaranteed to improve the variational objective at each step, or leave it unchanged. Our method is very simple and comes with negligible additional cost, and does not require as many iterations as simulated annealing. As discussed in the sequel, the asymptotic nature makes AVI similar in spirit to other small-variance asymptotic methods. Our experiments show significant performance gain over traditional anneal-

ing methods on Latent Dirichlet Allocation (LDA), Hidden Markov Model (HMM) and Stochastic Variational Inference (SVI).

We start with a brief discussion on variational inference in the next section. In Section III we both review the previous work on annealing and introduce our method. Section IV illustrates the benefits of the proposed method using a simple example. Finally, extensive experiments in Section V show the effectiveness of our method on real datasets.

II. Variational Inference

Given data X, parameters $\Theta = \{\theta_i\}$ for a model $p(X|\Theta)$, and prior $p(\Theta)$, the goal of Bayesian inference is to compute the posterior distribution $p(\Theta|X)$. This is usually intractable, often leading to approximation using mean-field variational inference [8]. In this approach, parameters of a factorized q-distribution, $q(\Theta) = \prod_i q(\theta_i)$, are tuned to minimize the Kullback-Leibler divergence

$$\mathrm{KL}(q||p) = \int q(\Theta) \ln \frac{q(\Theta)}{p(X|\Theta)} d\Theta. \tag{1}$$

Since this objective function cannot be obtained in closed form we instead maximize the surrogate evidence lower bound,

$$\mathcal{L} = \mathbb{E}_q[\ln p(X, \Theta)] + \mathbb{H}[q(\Theta)], \tag{2}$$

where $\mathbb{H}$ is the entropy function. For conjugate exponential family (CEF) models, optimization is often done over the natural parameters of each q-distribution, which we denote as $\boldsymbol{\lambda} = \{\lambda_i\}$.

A standard way to solve this is by gradient ascent

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \gamma_t \mathbf{M}_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}|_{\boldsymbol{\lambda}_t}, \tag{3}$$

where γ_t is a step size and $\mathbf{M}_t$ is a preconditioning matrix. In practice this is usually done using coordinate ascent for each λ_i separately, holding the others fixed. For CEF models, the gradient comes in a simplified form,

$$\nabla_{\lambda_i} \mathcal{L} = -\left(\frac{d^2 \ln q(\theta_i)}{d\lambda_i d\lambda_i^T} \right) (\mathbb{E}_q[\boldsymbol{t}] + \lambda_{0i} - \lambda_i). \tag{4}$$

The vector $\mathbb{E}_q[\boldsymbol{t}]$ is the expected sufficient statistics with respect to all q other than $q(\theta_i)$ and λ_{0i} is the prior. By setting $\gamma_t = 1$ and $\mathbf{M}_t$ to the inverse Fisher information of $q(\theta_i)$, the natural gradient results in the optimal coordinate update

$$\lambda_i \leftarrow \mathbb{E}_q[\boldsymbol{t}] + \lambda_{0i}. \tag{5}$$

While this is the optimal coordinate update for λ_i, iterating this coordinate ascent method over variables i will only converge to a local optimal solution. It has previously been noted that annealing can improve this local optimum, which we review below, along with our proposed annealing method.

III. Annealing and AVI

A. Deterministic Annealing

Before developing our method we briefly review deterministic annealing [11] which has been an effective technique for variational inference. The main idea here is to anneal the distribution by scaling the entropy term in the evidence lower bound. This encourages smoother distributions that have higher entropy and can help reduce the large number of modes. The modified, iteration-dependent objective is

$$\mathcal{L}_t = \mathbb{E}_q[\ln\, p(X, \Theta)] + T_t \mathbb{H}[q(\Theta)]. \tag{6}$$

Here T_t is the temperature variable which decays as the iterations increase. Therefore, in the early iterations the impact of the entropy is strong, and transitions to standard variational inference as $T_t \to 1$.

Taking the derivative of the lower bound with respect to λ_i and taking the natural gradient as before gives the optimal update

$$\lambda_i \leftarrow \frac{1}{T_t}(\mathbb{E}_q[\boldsymbol{t}] + \lambda_{0i}). \tag{7}$$

As is evident, deterministic annealing down-weights the amount of information in the posterior, thus increasing the entropy, but the modification is still determined by the data.

B. Simulated annealing

Our proposed method is based on simulating random perturbations of the deterministic variational updates, and is therefore a simulated annealing-type method. In the context of variational inference for a CEF model, the main idea of simulated annealing is to update the natural parameters as

$$\boldsymbol{\lambda}'_{t+1} \leftarrow \boldsymbol{\lambda}_t + \gamma_t \mathbf{M}_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}|_{\boldsymbol{\lambda}_t} + T_t \varepsilon_t \ , \tag{8}$$

where ε_t is a random noise vector controlled by the temperature variable $T_t \geq 0$. Here $T_t \to 0$ as $t \to \infty$. These updates are accepted according to the probability

$$\Pr(\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}'_{t+1}) = \min\left[1, \exp\left\{-\tfrac{\mathcal{L}(\boldsymbol{\lambda}_t) - \mathcal{L}(\boldsymbol{\lambda}'_{t+1})}{T_t}\right\}\right]. \tag{9}$$

Otherwise $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t$. In the early stages, $\boldsymbol{\lambda}_t$ is volatile enough to escape shallow optima. As the temperature T_t decreases (at an appropriate rate), this algorithm mimics the standard variational inference updates and converges. Again, this is usually performed separately over each λ_i.

Historically, noise enhancement was observed in physical systems [15] and simulated annealing was first used for discrete variables [10], [16], [17]. This was later extended to continuous random variables and analyzed in the context of continuous-time processes [18]. In [18] and [19] the authors showed that under certain conditions simulated annealing weakly converges to the global optimum of $\mathcal{L}$ as $T \to 0$ and $t \to \infty$, which gives theoretical justification for using this kind of algorithm. In [20] and [21], [22], discrete-time versions of this was developed that have the same convergence property. These ideas also found applications in the machine learning domain, such as Hamiltonian Monte Carlo [23], importance sampling [24], and stochastic gradient Langevin dynamics

(SGLD) [25], [26], where the temperature T_t is decreasing at a slower rate.

While simulated annealing is a popular choice for many optimization problems [2], it is not as well-suited for variational inference as used in machine learning. This is because variational inference is typically only run for a small number of iterations - contrary to MCMC - where the benefit of simulated annealing may not manifest itself. Furthermore, since the algorithm takes random steps there is no guarantee that we are improving the variational objective at every step. We argue that an algorithm should have such improvement guarantees when run for so few iterations.

C. Annealing with AVI

We derive our very simple annealing approach as a modification of simulated annealing. Using coordinate ascent in(8) over individual λ_i, we choose the values

$$\boldsymbol{M_i} = -\left(\frac{d^2 \ln q(\theta_i)}{d\lambda_i d\lambda_i^T}\right)^{-1},$$

$$T_t = \epsilon, \quad \gamma_t = 1, \quad \varepsilon_t = \frac{\rho_t}{\epsilon}(\eta_{t,i} - \mathbb{E}_q[\boldsymbol{t}] - \lambda_0). \tag{10}$$

We set ρ_t to be a step size that is shrinking to zero as t increases and discuss the random vector $\eta_{t,i}$ and $\epsilon \in \mathbb{R}_+$ shortly. The proposed update is therefore

$$\lambda_i' \leftarrow (1-\rho_t)(\mathbb{E}_q[\boldsymbol{t}] + \lambda_{0i}) + \rho_t \eta_{t,i}. \tag{11}$$

With respect to (7), note that $\frac{1}{T_t} \Leftrightarrow 1-\rho_t$; this first term therefore is similar to deterministic annealing. The second injects noise to further search the objective space. We propose and use a staircase cooling schedule for ρ_t, where ρ_t is kept constant for a fixed number of steps and then decreased in value. The value of ρ_t and the time window can be regarded as search radius and search duration, respectively. After a certain iteration I, $\rho_t = 0$ for $t > I$.

To determine whether we accept or reject λ_i', we propose the simple strategy

$$\text{accept} \quad \lambda_i' \quad \text{if} \quad \mathcal{L}(\lambda_i') > \mathcal{L}(\lambda_i^{(\text{old})}). \tag{12}$$

Algorithm 1 Asymptotically-annealed VI (AVI)

For CEF models where updates are in closed form.
Randomly initialize parameters: λ_{0i} for $q(\theta_i)$.
Obtain initial configuration:
$[\lambda_i^{\text{old}}\ ,\ \mathcal{L}^{\text{old}}] \leftarrow \text{VI}(\lambda_{0i})$.
for each $q(\theta_i)$ in iteration t **do**
  Set ρ_t according to staircase cooling.
  Generate noise $\eta_{t,i}$ for λ_i.
  Generate proposal: $\lambda_i^p \leftarrow (1-\rho_t)\lambda_i^{\text{old}} + \rho_t \eta_{t,i}$.
  Update: $[\lambda_i^{\text{new}}\ ,\ \mathcal{L}^{\text{new}}] \leftarrow \text{VI}(\lambda_i^p)$
  if $\mathcal{L}^{\text{new}} > \mathcal{L}^{\text{old}}$
    $\lambda_i^{\text{old}} \leftarrow \lambda_i^{\text{new}}$ and $\mathcal{L}^{\text{old}} \leftarrow \mathcal{L}^{\text{new}}$
  end if
end for
Note: $[\lambda^{(t+1)}\ ,\ \mathcal{L}^{(t+1)}] \leftarrow \text{VI}(\lambda^{(t)})$ reads in the current variational parameters, updates them and computes the lower bound for the new parameter setting.

Unlike simulated annealing, which assumes that many iterations will be done to fully explore the parameter space, this algorithm simply accepts if there's improvement or rejects otherwise. It is therefore guaranteed to monotonically improve the lower bound. Since we set $\rho_t = 0$ for $t > I$, convergence follows by standard rules [8]. As for $\eta_{t,i}$ *any* random variable that is in the parameter's support set can be used. We summarize the algorithm in Algorithm 1. Finally, note that in the pseudocode the injected noise is written in the form of convex combination due to our derivation; but practically noise can be added in any desired way.

a) Asymptotic connections: We can connect our proposed AVI technique to an asymptotic limit of simulated annealing. In particular, we notice that, as formulated in (10), in the limit as $\epsilon \to 0$ the setting of λ_i' in (11) remains unchanged because ϵ cancels in the product $T_t\varepsilon_t$. However, if we then accept/reject according to the simulated annealing procedure in (9), the probability of accepting is 0 or 1, depending on whether the new value improves the objective function. This is similar in spirit to small variance asymptotic approaches that have been derived for a variety of Bayesian models [27]–[29]. There, the motivation is partly to avoid having to run many iterations using a sampling algorithm; for example, models such as DP-mixtures can reduce to "nonparametric K-means" that are much faster to compute. Our proposed AVI has a similar advantage of requiring far fewer iterations than standard simulated annealing approaches, while still having annealing properties.

b) Relation to previous annealing work: We have previously discussed how this method can be thought of as building on deterministic simulated annealing. Comparing lower bound values corresponds to interacting states which is inspired by quantum annealing [14], but our strategy is to work with the M-step while quantum annealing concentrates on the E-step. Our usage of a staircase cooling is motivated by the analysis of [13]. Interestingly, when applied to SVI this algorithm can be regarded as a reversal of the SGLD method [25]; in SGLD, the gradient dominates at first and the injected noise dominates later, while the proposed method does the reverse.

IV. An Illustrative Example

We demonstrate our annealing approach over a commonly evaluated test model [11], [13]. In this model, the data is generated from a mixture of two one-dimensional Gaussians with means μ_1, μ_2, a shared standard deviation of σ, and cluster mixing vector α. We place a normal prior over the unknown means μ_k. Defining a factorized $q(\mu_k) = N(\hat{\mu_k}, \hat{\sigma}_k^2)$ for $k = 1, 2$, the variational updates are

$$\hat{\phi}_k^i = \frac{\alpha_k \exp\{-\frac{1}{2\sigma^2}(x_i - \hat{\mu}_k)^2 - \frac{1}{2}\frac{\hat{\sigma}_k^2}{\sigma^2}\}}{\sum_j \alpha_j \exp\{-\frac{1}{2\sigma^2}(x_i - \hat{\mu}_j)^2 - \frac{1}{2}\frac{\hat{\sigma}_j^2}{\sigma^2}\}} \tag{13}$$

$$\hat{\mu}_k = \frac{\sigma^2\big(\sum_i \hat{\phi}_i^k\big)}{1+\sigma^2\big(\sum_i \hat{\phi}_i^k\big)}\Big(\frac{\sum_i \hat{\phi}_i^k x_i}{\sum_i \hat{\phi}_i^k}\Big), \tag{14}$$

$$\hat{\sigma}_k^2 = \frac{\sigma^2}{1+\sigma^2\big(\sum_i \hat{\phi}_i^k\big)} \tag{15}$$

In our annealing approach, we update $\hat{\mu}_k$ by first forming the update above, $\hat{\mu}'_k$, and averaging with a random initialization, η, for $\hat{\mu}_k$ such that $\hat{\mu}_k \leftarrow (1-\rho)\hat{\mu}'_k + \rho\eta$ (omitting iteration subscripts). We note that this is in contrast with deterministic annealing, which inflates $\hat{\sigma}^2_k$ and $\hat{\phi}^i_k$ and leaves $\hat{\mu}_k$ untouched. We synthesize data using identical settings as in [11]. We plot the objective function restricted to the mean parameters of each q distribution in Figure 1.

In Figure 2, we compare variational inference (VI) with deterministic annealing (DAVI), simulated annealing (SAVI), and the proposed AVI. We use 250,000 initializations along a uniform grid. The plots show the probability of global optimal convergence as a function of starting point (blue = 0 and red = 1). (We use kernel smoothing to make these plots for SAVI and AVI. For VI and DAVI we simply show the binary values as these approaches are deterministic.) We also show the overall success probability above each plot. For DAVI we use the cooling schedule of [11]. AVI and SAVI use the same cooling scales and both stochastic annealing algorithms start from the same temperature; making them directly comparable. We set these schedules to all have the same search radius so that comparison with DAVI is also meaningful.

We notice that DAVI and VI have linear boundaries, being deterministic, while the stochasticity of AVI and SAVI allow for probabilities to wrap into the local optimal region. While stretching out SAVI over many more iterations and with a slower cooling schedule produced better results, this removes the speed advantage of VI over MCMC inference—for the large number of iterations required for SAVI to theoretically work well, we argue that AVI is preferable. These results show that our asymptotic approach to simulated annealing has advantages. We also observe that, for the same "amount" of cooling, AVI can achieve better convergence than DAVI.

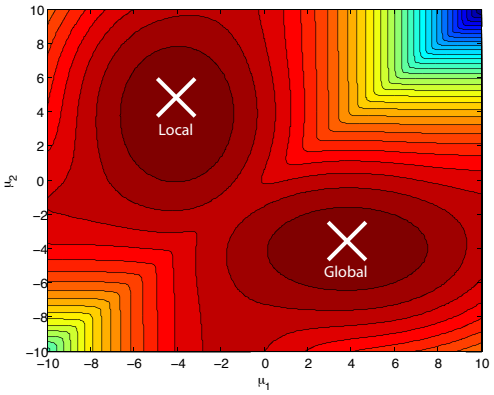


Fig. 1. Contoured heatmap of the variational objective function used in our toy experiment showing the two optima.

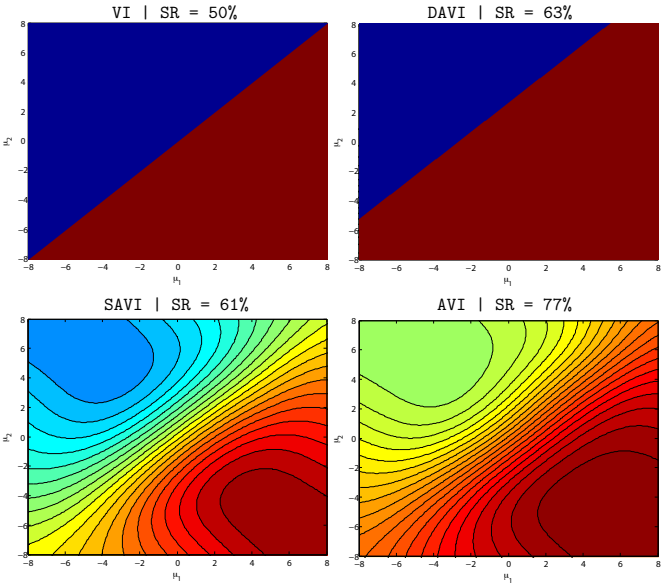


Fig. 2. Heatmaps showing the probability of converging to the global optimum given the starting point; red = 1 and blue = 0. Results are shown for variational inference (VI), deterministic annealing (DAVI), simulated annealing (SAVI) and asymptotic simulated annealing (AVI). Also shown is the overall probability of converging to a global optimum given a starting point picked uniformly at random.

V. Experiments

In this section we perform a number of experiments to compare AVI with other methods. For batch variational inference we evaluate using two models: Latent Dirichlet allocation (LDA) [30] and the discrete hidden Markov model (HMM) [31]. We compare our method to deterministic annealing (DAVI) and simulated annealing (SAVI). We also test our algorithm in the stochastic inference setting, comparing it to the recently proposed variational tempering (VT) [13].

A. Latent Dirichlet Allocation

We begin our comparisons with LDA. We consider the three corpora indicated in Table I. The model variables for a K-topic LDA model are $\{\beta_{1:K}, \pi_{1:D}, \boldsymbol{c}_d\}$ where D is the number of documents. The vector π_d gives a distribution on topics in the set $\boldsymbol{\beta}$ for document d. Each topic β_k is a distribution on V vocabulary words. The vector $\boldsymbol{c}_d$ indicates the allocation of each word in document d to a topic.

We use a mean-field posterior approximation

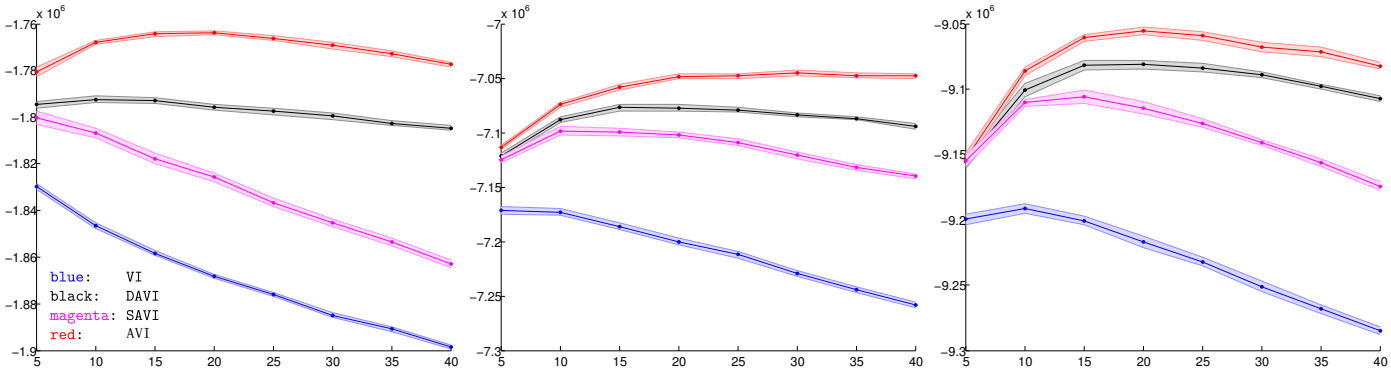
$$q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{c}) = \Big[\prod_k q(\beta_k)\Big]\Big[\prod_d q(\pi_d)\Big]\Big[\prod_{d,n} q(c_{d,n})\Big].$$

Here q distributions on β_k and π_d are Dirichlet and $c_{d,n}$ are multinomial. We set the Dirichlet prior parameter of π_d to $1/K$ and the Dirichlet prior parameter of β_k to $100/V$ and initialize all Dirichlet q distributions from a scaled uniform Dirichlet random vector.

When no annealing is present, variational parameter update corresponds to summing expected counts over all words and documents. The updates for $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ have similar structures. Focusing on topic vectors, the update for the variational parameter λ_k of β_k is

$$\lambda_k = \textstyle\sum_{d,n} \phi_{d,n}(k) w_{d,n} + \lambda_0 \ . \tag{16}$$

Here the variables $\phi_{d,n}$ denote the allocation probability vector across topics of nth word in dth document. The indicator vector $w_{d,n}$ (the data) corresponds to this word value.



(a) arXiv (b) Huffington Post (c) New York Times

Fig. 3. Comparison of the variational objective as a function of topics for three different algorithms: standard variational inference without annealing (VI), deterministic annealing (DAVI), simulated annealing (SAVI) and AVI. We see that in all cases AVI clearly outperforms the others algorithms. We also observe that annealing can provide more accurate information for model selection over the number of topics.

As shown in Section III, AVI is applied without modifying the collecting of these statistics. We just pick a suitable noise distribution to simulate the Markov chain. To update the topics $\boldsymbol{\beta}$ we choose a scaled Dirichlet noise vector, $\eta_i \sim \text{Dir}(1, \ldots, 1)$, and obtain a proposal of the form

$$\lambda_k^p = (1-\rho)(\textstyle\sum_{d,n} \phi_{d,n}(k) w_{d,n} + \lambda_0) + \rho\gamma\eta_k, \qquad (17)$$

where again we suppress iteration number. We use γ to scale the probability vector η_k, which ensures the noise is not too weak.

The remaining simulation set-up is as follows: Deterministic and simulated annealing uses the cooling schedule $T_t = 1 + 2 \times 0.7^t$. These values give significant improvement over the baseline variational inference algorithm and was found to give good representative results compared with other strategies. For AVI we set $\rho_0 = 0.3$ and follow a staircase cooling where $\rho_t \in \{0.3, 0.2, 0.1\}$. This way the temperature ranges of all algorithms are matched and we get fair comparison. All algorithms use 100 iterations and for DAVI/SAVI/AVI we turn off modifications after 75 steps. This ensured the convergence of parameters to their final values. We use these settings for all LDA experiments.

In Figure 3 we show plots of the final value of the variational objective as a function of K averaged over multiple runs. Firstly we notice that there is a clear benefit of employing annealing in general. AVI, in turn, is capable of improving the lower bound quite significantly and consistently for all experiments. Annealing also helps with model selection, i.e. finding a good value for K. For most of the cases VI simply gives a decreasing lower bound which is not very informative. AVI gives the best lower bound values and we see that in most cases its optimal K value disagrees with DAVI/SAVI.

The fact that AVI performs significantly better than SAVI shows the two algorithms are quite different in nature. For this and other models we tried, DAVI works better than SAVI as well when both are restricted to a small number of iterations and for this reason simulated annealing has not found much use in variational inference applications. In high dimensional problems such as LDA the traditional simulated annealing would naturally require a large number of iterations to give good results. We therefore think this experiment shows the significant advantage of AVI in settings where the number of iterations is restricted to be small. In Table II we show the probability of accepting the proposal (i.e., improving the variational objective) after adding noise. We see that most of the proposals are accepted because the information from the data in the update provides enough improvement that the random noise does not take the parameter to a worse location in the space.

B. Hidden Markov Model

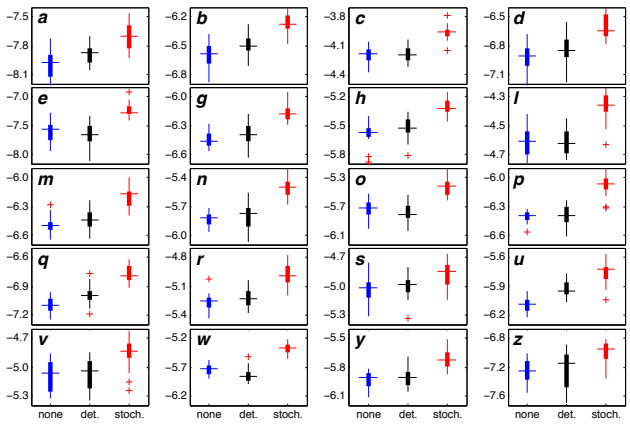
Next, we consider the discrete hidden Markov model with K-states. Similar to LDA, this model has proved useful for many applications [32]. Here we present our method on a character trajectories dataset from the UCI Machine Learning Repository, which consists of sequences of spatial locations of a pen as it is used to write one of 20 different characters. In

TABLE I
STATISTICS OF THE THREE CORPORA USED IN OUR ANNEALING EXPERIMENTS FOR LDA.

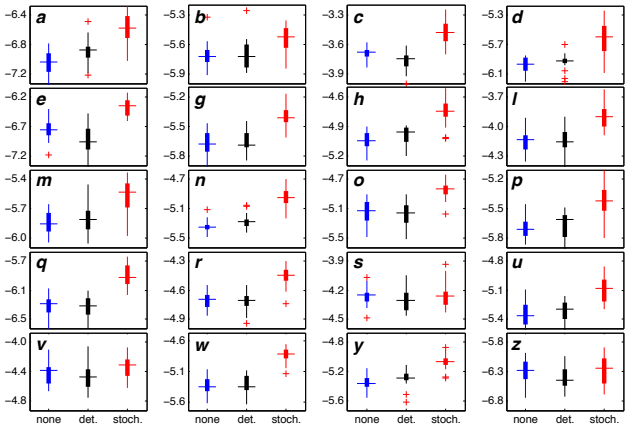
	ArXiv	HuffPost	NYT
# docs	3.8K	4K	8.4K
# vocab	5K	6.3K	3.0K
# tokens	234K	906K	1.2M
# word/doc	62	226	143

TABLE II
ACCEPTANCE PROBABILITY FOR AVI AT EACH STAIRCASE LEVEL FOR ARXIV. AVERAGED OVER TEN RUNS.

#topics	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
15	1	0.97	0.90
20	1	0.99	0.88
25	1	0.99	0.92



(a) 5-state hidden Markov model



(b) 10-state hidden Markov model

Fig. 4. Box plots for variational objective function ($\times 10^4$) for each character.The color codes are blue/VI, black/DAVI, and red/AVI. Our algorithm consistently gives the best improvements and for most of the characters, the improvement margin is large.

this dataset, there are 2,858 sequences in total from which we hold out 200 for testing (ten for each character). We quantized the 3-dimensional sequences using a codebook of size 500 learned using K-means.

For this model the model variables are $\{\pi, A, B\}$, with hidden state sequences $\boldsymbol{s}_n$ for each observed sequence. The variable π is an initial state distribution, A is the Markov transition matrix and B gives the emission probabilities over a discrete set. K is the number of hidden states. Once again we consider a mean-field factorization of form

$$q(\pi, A, B, \boldsymbol{s}) = q(\pi) \prod_{k=1}^{K} q(A_{k,:}) q(B_{k,:}) \prod_n q(\boldsymbol{s}_n)$$

where the individual factors on π, A and B are Dirichlet and a discrete distribution on each sequence $\boldsymbol{s}_n$. For the priors on A and π we set the Dirichlet parameter to $1/K$. For the priors on B we set the Dirichlet parameter to $10/V$, where V is codebook size. As with LDA, we initialize all q distributions by scaling up a uniform Dirichlet random vector.

Also similar to LDA, the variational updates for the factors have the general form [31]

$$\lambda_k = \sum_n \sum_m \phi_k^{nm} + \lambda_0, \tag{18}$$

where λ_0 is a prior and ϕ_k^{nm} is a probability relating to the mth emission in sequence n and state k. Similar to previous section we modify this update to obtain the proposal

$$\lambda_k^p = (1-\rho)(\sum_n \sum_m \phi_k^{nm} + \lambda_0) + \rho\gamma\eta_k, \tag{19}$$

where we used additive scaled Dirichlet noise. Again we suppress iteration index t.

Figure 4 shows box plots of variational lower bound for a 5-state and 10-state HMM, where we compare with variational inference and deterministic annealing. We see that AVI provides a major improvement over VI and DAVI.[1] The improvement provided by DAVI is not significant for a number of cases. As is clear, AVI provides a significant improvement over deterministic annealing, which only deforms the objective landscape, but does not explore over the variational parameter space. While AVI deforms the landscape in a similar way by pre-multiplying the updates by $(1-\rho)$, the additional exploration provided by the stochastic part η_k is seen to be a significant advantage.

[1]SAVI results are not better than DAVI and are omitted to reduce clutter in figures.

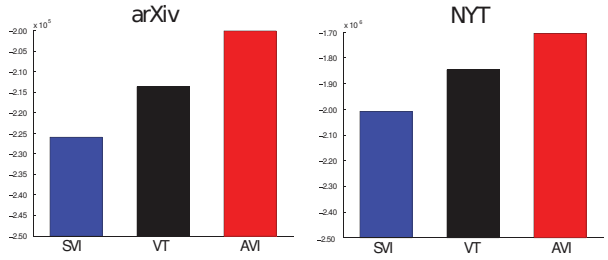


Fig. 5. Bar charts of variational objective function for arXiv and big NYT datasets. Our algorithm improves significantly upon the state-of-the-art variational tempering technique.

C. Stochastic Variational Inference

Our last experiment uses stochastic variational inference, which allows for fast variational inference over very large datasets [7]. This method randomly subsamples mini-batches to obtain unbiased estimates of the gradients, and partially updates the global variables using convex combinations controlled by a step size. Unlike the previous settings we considered, here the full objective function we are optimizing is unavailable as it requires computing a lower bound over a large collection of samples, which would compromise the scalability provided by stochastic gradients. Therefore convergence is typically monitored over a small independent validation set, which we also adopt.

We apply the AVI to the SVI updates and use the validation set lower bound for comparisons. AVI does not introduce additional complexity to the SVI framework, since the updates

are simply weighted averages of the true updates with random noise. We apply AVI only to the global variable in this case. We again use LDA as the model, applying it to a New York Times corpus containing roughly 1.8 million documents and the small arXiv dataset shown in Table I. The update and proposal parameters are similar to Eqs. (16) and (17).

We compare our method to variational tempering (VT) [13] which is a state-of-the-art method that generalizes deterministic annealing to SVI; in particular we use local variational tempering as it gives good results on LDA. The remaining simulation settings are as follows: For arXiv we hold out 358 documents on which we evaluate the variational objective. We set the number of topics to $K = 50$ and batch size to $B = 100$. The step size as a function of iteration is set to $(50 + t)^{-0.51}$. VT and AVI are activated for the first 300 iterations following a staircase cooling schedule, and then are turned off to ensure convergence. For the large New York Times corpus we hold out 625 documents to evaluate the variational objective, set $K = 50$, $B = 500$, and use the same cooling as in arXiv.

Figure 5 shows the variational objective averaged over multiple runs (error bars were very small so are omitted). For both cases AVI yields a clear improvement over VT, just as VT clearly improves the base SVI algorithm. In addition, it requires less computation since VT infers an additional M-dimensional latent temperature variable per data point. Therefore we gain better performance for less computation.

VI. Conclusion

We have introduced a simple yet effective annealing technique for variational inference, which we call asymptotic simulated annealing. We showed how this technique can be built into the existing variational inference procedure with minimal modification, making it easy to use for practitioners. We demonstrated the benefits of this algorithm on highly non-convex problems, such as LDA and the HMM. We also showed the effectiveness on stochastic variational inference, improving the state-of-the-art variational tempering, an extension of deterministic annealing to SVI.

Acknowledgments

We thank the anonymous reviewers for detailed comments and suggestions.

References

[1] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

[2] Christian Robert and George Casella, *Monte Carlo statistical methods*, Springer Science & Business Media, 2013.

[3] Darryl D. Lin and Teng J. Lim, "A variational free energy minimization interpretation of multiuser detection in CDMA.," in *Global Telecommunications Conference (GLOBECOM)*, 2005.

[4] Darryl D. Lin and Teng J. Lim, "A variational inference framework for soft-in soft-out detection in multiple-access channels," *IEEE Transactions on Information Theory*, 2009.

[5] Feng Li, Zongben Xu, and Shihua Zhu, "Variational inference based data detection for OFDM systems with imperfect channel estimation," *IEEE Transactions on Vehicular Technology*, 2013.

[6] Gunvor Kirkelund, Carles Manchon, Lars Christensen, Erwin Riegler, and Henri Fleury, "Variational message-passing for joint channel estimation and decoding in MIMO-OFDM," in *Global Telecommunications Conference (GLOBECOM)*, 2010.

[7] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.

[8] Martin J. Wainwright and Michael I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 2008.

[9] James Kennedy, *Particle Swarm Optimization*, pp. 760–766, Springer US, Boston, MA, 2010.

[10] Scott Kirkpatrick, Daniel Gelatt, and Mario Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[11] Kentaro Katahira, Kazuho Watanabe, and Masato Okada, "Deterministic annealing variant of variational bayes method," *Journal of Physics: Conference Series*, 2008.

[12] Ryo Yoshida and Mike West, "Bayesian learning in sparse graphical factor models via variational mean-field annealing," *Journal of Machine Learning Research*, pp. 1771–1798, 2010.

[13] Stephan Mandt, James McInerey, Farhan Abrol, Rajesh Ranganath, and David Blei, "Variational tempering," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[14] Issei Sato, Kenichi Kurihara, Shu Tanaka, Hiroshi Nakagawa, and Seiji Miyashita, "Quantum annealing for variational Bayes inference," in *Uncertainty in Artificial Intelligence (UAI)*, 2009.

[15] Roberta Benzi, Giorgio Parisi, Alfonso Sutera, and Angelo Vulpiani, "Stochastic resonance in climatic change," *Tellus*, 1982.

[16] Vladimir Cerny, "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm," *Journal of Optimization Theory and Applications*, 1985.

[17] Stuart Geman and Donald Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[18] Stuart Geman and Chii-Ruey Hwang, "Diffusions for global optimization," *SIAM Journal on Control and Optimization*, 1986.

[19] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn J. Sheu, "Diffusion for global optimization in $\mathbb{R}^n$," *SIAM Journal on Control and Optimization*, 1987.

[20] Harold Kushner, "Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo," *SIAM Journal on Applied Mathematics*, 1987.

[21] Saul B. Gelfand and Sanjoy K. Mitter, "Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$," *SIAM Journal on Control and Optimization*, 1991.

[22] Saul B. Gelfand and Sanjoy K. Mitter, "Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$," *SIAM Journal on Control and Optimization*, 1993.

[23] Radford M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, Eds. CRC Press, 2010.

[24] Radford M. Neal, "Annealed importance sampling," *Statistics and Computing*, 2001.

[25] Max Welling and Yee Whye Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *International Conference on Machine Learning (ICML)*, 2011.

[26] Sungjin Anh, Korattikara Anoop, and Max Welling, "Bayesian posterior sampling via stochastic gradient Fisher scoring," in *International Conference on Machine Learning (ICML)*, 2012.

[27] Brian Kulis and Michael I. Jordan, "Revisiting k-means: New algorithms via Bayesian nonparametrics," in *International Conference on Machine Learning (ICML)*, 2012.

[28] Ke Jiang, Brian Kulis, and Michael I. Jordan, "Small-variance asymptotics for exponential family Dirichlet process mixture models," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 3158–3166.

[29] Tamara Broderick, Brian Kulis, and Michael I. Jordan, "MAD-Bayes: MAP-based asymptotic derivations from Bayes," in *International Conference on Machine Learning (ICML)*, 2013.

[30] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, pp. 993–1022, Mar. 2003.

[31] Matthew J. Beal, "Variational algorithms for approximate Bayesian inference," *Ph. D. Thesis, University College London*, 2003.

[32] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 1989.