OVERVIEW OF LARGE SCALE OPTIMIZATION FOR DISCRIMINATIVE TRAINING IN SPEECH RECOGNITION

Dimitri Kanevsky[1], *Georg Heigold*[2], *Stephen Wright*[3], *Hermann Ney*[4]

[1]IBM T. J. Watson Research Center, Yorktown Heights, USA
[2]Google, Mountain View, USA
[3]University of Wisconsin, Madison, US
[4]Lehrstuhl für Informatik, RWTH Aachen, Germany
[1]kanevsky@us.ibm.com, [2]heigold@google.com, [3]swright@cs.wisc.edu, [4]ney@informatik.rwth-aachen.de

ABSTRACT

Over the past few decades, a variety of specialized approaches have been proposed to solve large problems in speech recognition. Conventional optimization techniques have not been widely applied, because the problems do not readily admit an objective for evaluating a given set of parameters and because of the large number of parameters. This situation is changing, due to recent developments in algorithmic optimization. In this paper, we review the specialized algorithms, including methods derived from the extended Baum-Welch (EBW) approach, Rprop, and GIS. We discuss optimization frameworks that could also potentially be applied, and outline some connections between the optimization methods and existing specialized methods.

Index Terms— GIS, auxiliary function, Rprop, EBW

1. INTRODUCTION

Over the past few decades, a variety of specialized approaches have been proposed for optimizing generative models for conditional probability densities. Conventional optimization techniques have not been widely applied, because the problems do not readily admit an objective for evaluating a given set of parameters and because of the large number of parameters. One of the approaches to optimize conditional likelihoods is a family of algorithms, called extended Baum-Welch (EBW) ([5], [6], [7]), that is the current state-of-the-art in speech processing with hidden Markov models. In this paper, we review the specialized algorithms, including methods derived from the extended Baum-Welch (EBW) approach, Rprop, and GIS. We discuss optimization frameworks that could also potentially be applied, and outline some connections between the optimization methods and existing specialized methods.

Section 2 contains a brief overview of EBW methods. In Section 3 we describe $\mathcal{A}$-functions, which form the basis of several algorithms, and outline some techniques for constructing them. Section 4 describes an explicit auxiliary function for discriminative training of Gaussian mixture models of HMMs, while Section 5 describes gradient steepness metrics associated with $\mathcal{A}$-functions. The Rprop approach is discussed in Section 6, while Section 7 discusses general optimization methodologies and speculates about their possible application to speech recognition problems.

2. EBW

Let $f(\xi)$ be some differentiable function in variables $\xi = \{\xi_t(\theta)\}$ where θ is some parameter. Let $c_t(\theta) = \xi_t(\theta)\frac{\partial f(\xi)}{\partial \xi_t}$. Consider the case of multidimensional multivariate Gaussian densities:

$$\xi_t = \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} e^{-1/2(x_t-\mu)^T\Sigma^{-1}(y_i-\mu)} \tag{1}$$

where $X_1^T = \{x_t \in \mathcal{X}\}, t = 1,..T$ is a training sample. Then the EBW updates for parameters $\theta = \{\mu, \Sigma\}$ are defined as following.

$$\hat{\mu} = \frac{\sum_t c_t(\theta)x_t + D\mu}{\sum_t c_t(\theta) + D} \tag{2}$$

$$\hat{\Sigma} = \frac{\sum_t c_t(\theta)x_t x_t^T + C(\mu\mu^T + \Sigma)}{\sum_t c_t(\theta) + D} - \hat{\mu}\hat{\mu}^T \tag{3}$$

The proof that the transformations (2) yield growth in the function f, for sufficiently large D, can be seen from Section 3.2.

We summarize a few known results on EBW. It was shown in [11] that a unified objective function that includes as special cases two major approaches in discriminative training — Maximum Mutual Information (MMI) and Minimum Classification Error (MCE) — can be optimized using EBW updates (2). In [12], it was shown that preventing update models to be too far from initial models in the EBW update formula allows additional improvements in the recognition accuracy. It was shown in [13] that EBW for a MMI objective function comes from a regularization that is based on Kulback-Leibler (KL) divergence between two probability distributions.

3. GENERALIZATIONS FOR EBW

In this section following [8] we show that EBW process is a special case of transformations that involve various generalizations of auxiliary functions.

3.1. $\mathcal{A}$-functions

Let $f(x) : \mathcal{U} \subset \mathbb{R}^n \to \mathbb{R}$ be a real valued differentiable function in an open subset $\mathcal{U}$. The function $\mathbf{A}_f = \mathbf{A}_f(x, y) : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ is called an $\mathcal{A}$-function for f if it is twice differentiable in $x \in \mathcal{U}$ for each $y \in \mathcal{U}$ and if the following properties hold.

1 $\mathbf{A}_f(x, y)$ is a strictly convex or strictly concave function of x for any $y \in \mathcal{U}$. (Recall that a sufficient condition for a twice

differentiable function to be strictly concave or convex over some domain is that its Hessian function is positive definite or negative definite in the domain, respectively.)

2 Hyperplanes tangent to manifolds defined by $z = g_y(x) := \mathbf{A}_f(x, y)$ and $z = f(x)$ at any $x = y \in \mathcal{U}$ are parallel to each other, that is,

$$\nabla_x \mathbf{A}_f(x, y)|_{x=y} = \nabla_x f(x). \tag{4}$$

If in addition $\mathbf{A}_f$ is concave and lower bounded — that is, $\mathbf{A}_f(x, y) \geq f(x)$ for any $x, y \in \mathcal{U}$ — then $\mathbf{A}_f$ is called an auxiliary function.

3.2. LSAF

The $\mathcal{A}$-function for f define above can be used to deduce iterative methods for maximizing f.

Let θ_0 be some point in $\mathcal{U}$ and $\mathcal{U} \ni \tilde{\theta}_0$ be a solution of the nonlinear algebraic equation $\nabla_\theta \mathbf{A}_f(\theta, \theta_0)|_{\theta=\tilde{\theta}_0} = 0$. ($\tilde{\theta}_0$ is the minimizer of $\mathbf{A}_f(\theta, \theta_0)$ if $\mathbf{A}_f$ is convex and the maximizer if $\mathbf{A}_f$ is concave.) We have the following *growth statement* concerning small steps along the direction defined by $\tilde{\theta}_0 - \theta_0$: Defining

$$\theta(\alpha) := \alpha\tilde{\theta}_0 + (1 - \alpha)\theta_0, \tag{5}$$

we have for sufficiently small $|\alpha| \neq 0$ that $f(\theta(\alpha)) > f(\theta_0)$ where $\alpha > 0$ if $\mathbf{A}_f(\theta, \theta_0)$ concave and $\alpha < 0$ if $\mathbf{A}_f(\theta, \theta_0)$ convex. If $\mathbf{A}_f$ is an auxiliary function, then the growth property for f still holds for $|\alpha| = 1$ in (5). In the latter case, computation of $\tilde{\theta}_0$ and $\theta(1)$ represent, respectively, the E and M steps in Expectation-Maximization (EM) algorithm. An example of an $\mathcal{A}$-function is the following (associative) function

$$\sum_t c_t(\theta_0) \log \xi_t(\theta) \tag{6}$$

where $\xi_t(\theta)$ are Gaussian, Poisson, Gamma densities, or exponential family of densities. (Exponential family of densities are defined as $\frac{\exp\{\theta^T \phi(x_t)\}}{Z(\theta)}$, where the vector x_t is a base observation, the vector function ϕ characterizes the exponential family, and Z is the partition function).

4. AN EXPLICIT AUXILIARY FUNCTION FOR DISCRIMINATIVE GAUSSIAN MODELS OF HMMS

This section discusses construction of an auxiliary function for the discriminative training of Gaussian mixture models. Existing auxiliary functions such as Expectation-Maximization (EM) and Generalized Iterative Scaling (GIS) do not directly apply to this problem, because EM is for *generative* training of Gaussian models, while GIS is for MMI training of log-linear models *without* hidden variables.

An auxiliary function for discriminative Gaussian models can be derived by combining the auxiliary functions associated with EM and GIS [1, 2, Section 6.4]. More precisely, apply the auxiliary function associated with EM to the discriminative training objective for the Gaussian models. This step basically "eliminates" the hidden variables. We could stop here and use numerical optimization techniques to solve the M-step. Instead we rewrite the conditional probability induced by the Gaussian model as a log-linear model. The feature functions are of the type

$$f_{p\sigma d}(x, s) = \delta_{s\sigma} \cdot (x_{td})^{(p)}, \quad p \in \{0, 1, 2\}$$

where σ, s denote the mixture and d refers to the feature component. This second step does not add any flexibility to the model; the log-linear model can be converted to a valid Gaussian model after optimization [3]. This reparameterization leads to a linear combination of training objectives for log-linear models, so GIS applies. Optimization of the resulting auxiliary function leads to GIS-like update rules:

$$\theta_{psd}^{t+1} = \theta_{psd}^t + \frac{1}{F} \log \left(\frac{\sum_t c_{ts}^{(\text{num})} (x_{td})^p}{\sum_t c_{ts}^{(\text{den})} (x_{td})^p} \right)$$

with the numerator and denominator occupancies $c_{ts}^{(\text{num})}$ and $c_{ts}^{(\text{den})}$. Non-negative feature functions that sum up to the feature count $F = \max_{t,s} \sum_i f_i(x_t, s)$. (a requirement for GIS), are assumed without loss of generality. Like the iteration constant D for EBW, the feature count controls the convergence speed. In contrast to the iteration constant D in (2), the feature count F can depend on the training data. This does not affect the constructivity of the auxiliary function as long as the quantity can be explicitly computed before training.

This auxiliary function can be extended to other training objectives in the rational form such as Minimum Phone Error (MPE) [7] and to HMMs. In case of HMMs, the feature count scales with the number of frames in the sentence, which can slow down the convergence speed considerably.

5. GRADIENT STEEPNESS METRICS ASSOCIATED WITH $\mathcal{A}$-FUNCTIONS

The purpose of this section is to analyze how different gradient techniques associated with $\mathcal{A}$-functions are related between themselves. Specifically, one can interpret parameter update rules as aiming directly to improve recognition accuracy, that is, aiming to maximize the objective f.

In notation of the section of Section 2, let

$$\mathbf{A}_f(\theta, \theta_0) = \mathbf{B}(\{\xi_t(\theta)\}, \{\xi_t(\theta_0)\})$$

be $\mathcal{A}$-function for f. Consider the following gradient-ascent step:

$$\theta(\alpha) = \theta + \alpha \nabla_\theta \mathbf{A}_f(\theta, \bar{\theta})|_{\bar{\theta}=\theta}. \tag{7}$$

Let us also recall the update of parameters from Section 3.2, that is,

$$\hat{\theta}(\hat{\alpha}) = \theta + \hat{\alpha}(\tilde{\theta} - \theta) \tag{8}$$

where $\tilde{\theta}$ is a solution of

$$\nabla_{\bar{\theta}} \mathbf{A}_f(\bar{\theta}, \theta)|_{\bar{\theta}=\tilde{\theta}} = 0. \tag{9}$$

We say that (8) and (7) belong to the same family of solutions if for any sufficiently small α there exist $\hat{\alpha}$ such that

$$|\hat{\theta}(\hat{\alpha}) - \theta(\alpha)| < O(\alpha^2). \tag{10}$$

In other words, gradient-ascent and $\mathcal{A}$-function-based updates are the same up to first order. It can be shown that (10) holds in a case of diagonal Gaussian densities when $\mathbf{A}_f(\theta, \theta_0)$ is chosen to be the function (6). It was shown in [9] that EBW updates belong to the same family as updates via (8) $\mathcal{A}$-function. Thus, EBW updates and gradient descent updates for diagonal Gaussian densities belong to the same family. (This equivalence of EBW and gradient descent techniques was stated in [11].)

If $\mathbf{A}_f$ used to derive (8) satisifes the growth property, then in the linearization defined by

$$f(\hat{\theta}(\alpha)) - f(\theta) = T(\theta, \tilde{\theta}) * \alpha + O(\alpha^2) \tag{11}$$

we have

$$T(\theta, \tilde{\theta}) \geq 0. \tag{12}$$

For Gaussian densities, the paper [10] gives the exact expression for the term $T(\theta, \tilde{\theta})$ ($\theta = (\mu, \sigma)$, $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma})$), which is proportional to weighed sum of Euclidean distances $(\tilde{\mu} - \mu)^2$ and $(\tilde{\sigma}^2 - \sigma^2)$, where

$$\tilde{\mu} = \frac{\sum c_t(\mu, \sigma) x_t}{\sum c_t(\mu, \sigma)}, \quad \tilde{\sigma}^2 = \frac{\sum c_t(\mu, \sigma) x_t^2}{\sum c_t(\mu, \sigma)}$$

are solutions of (9) ([9]). One can introduce metrics $T(\theta, \tilde{\theta})$ in advance (for example as Kullback-Leibler distance between densities $\xi_t(\tilde{\theta})$ and $\xi_t(\theta)$). Following (8), one can define a recursion

$$\hat{\theta}(\hat{\alpha}) = \theta + \hat{\alpha} .* (\tilde{\theta} - \theta) \tag{13}$$

where $\hat{\alpha}$ is a vector and $.*$ is an element-wise product defined in such a way that (11) still holds. (Examples of such updates are given in [9] for Gaussian distributions.) Another way to change metrics is to make them proportional to accuracy measures (for example, frame or phonetic accuracy). Specifically, one can show that MPE method produces scaling of c_t coefficients that induces scaling on a metric T.

6. RPROP

Rprop, short for resilient backpropagation, is a gradient-based, batch update algorithm that uses adaptive step sizes. It was originally introduced for training of multilayer feedforward networks [19], but recent reports indicate its successful deployment in speech recognition [20, 4].

Rprop only uses the sign of the partial derivatives of the training objective for the parameter update

$$x_i^t = x_i^{t-1} + \mathsf{sign}\left(\frac{\partial f(x^t)}{\partial x_i}\right) \Delta_i^t.$$

There is a separate step size $\Delta_i^t \geq 0$, for each parameter x_i^t, updated independently at each iteration according to a simple heuristic. If the sign of the partial derivative changed over the last iteration, the step size is reduced by the positive factor $\eta^- < 1$. If the partial derivative kept the same sign, the step size is increased by the factor $\eta^+ > 1$. That is, we have

$$\Delta_i^t = \begin{cases} \eta^+ \Delta_i^{t-1}, & \text{if } \frac{\partial f(x^{t-1})}{\partial x_i} \frac{\partial f(x^t)}{\partial x_i} > 0 \\ \eta^- \Delta_i^{t-1}, & \text{if } \frac{\partial f(x^{t-1})}{\partial x_i} \frac{\partial f(x^t)}{\partial x_i} < 0 \\ 0, & \text{otherwise.} \end{cases}$$

The factors η^+ and η^- are set empirically; values that work well in practice are $\eta^+ = 1.2$ and $\eta^- = 0.5$. A fixed value $\Delta_i^0 := \Delta$ is chosen for the initial step size in each component. The parameter constraints for Gaussian mixture models such as the normalization of the mixture weights are re-imposed after each iteration.

We compare the performance of Rprop with EBW on two large-vocabulary continuous speech recognition tasks: European Parliament Plenary Sessions (EPPS) English from the TC-STAR project and Mandarin Broadcasts from the GALE project. In both cases, the standard RWTH setup is used. (See [4] for more details on the tasks and the specific setups.) Comparative results for EBW and Rprop are shown in Table 1 for Minimum Phone Error (MPE) training. The Maximum Likelihood (ML) baseline is added for comparison.

The two optimization algorithms achieve similar error rates on these two tasks, and both improve on ML. EBW converges in around ten iterations, typical of Gaussian mixture models with globally pooled variances. Rprop takes roughly the same number of iterations to converge for the conservative, default initial step size. For more aggressive initial step sizes, the training speeds up considerably, although at the risk of reduced convergence stability.

Table 1. Comparison of EBW and Rprop for discriminative training, word error rate (WER).

Task	Criterion	Optimization	WER [%]	
			Eval06	Eval07
EPPS English	ML	EM	10.8	12.0
	MPE	EBW	10.2	11.5
		Rprop	10.3	11.5
Mandarin Broadcasts	ML	EM	17.9	11.9
	MPE	EBW	17.0	11.2
		Rprop	16.5	11.1

7. OPTIMIZATION ALGORITHMS

The frameworks and algorithms described above, which have been developed mainly in the speech processing and statistics communities, have many connections to theory and algorithms for general optimization problems. These connections open up the possibility of applying many other recently developed optimization techniques to problems in speech recognition, including algorithms for sparse optimization and regularized logistic regression. We discuss some of these connections and mention a few related optimization approaches that may be adaptable to speech processing.

Approximation techniques similar to auxiliary functions have been developed recently in other contexts, for example regularized regression and compressed sensing. A simple function that satisfies the property (4) is

$$\mathbf{A}_f(x, y) = \nabla f(y)^T (x - y) - \frac{1}{2\alpha} \|x - y\|_2^2, \tag{14}$$

for some $\alpha > 0$ that plays the role of a line-search or trust-region parameter. The unconstrained minimum of (14) yields a steepest-ascent step. (The relationship between steepest-ascent and model-based steps was noted for several cases in Section 5.) If f is to be minimized over a convex set, (14) can be minimized over the same set to yield a gradient projection approach. In either case, the parameter α can be adjusted as needed to produce an improvement in f.

The approach based on (14) can be extended further to regularized optimization problems of the form

$$\max\ f(x) + h(x),$$

where $h(x)$ is a (typically nonsmooth but simple) function that is used to induce some desired structure in the solution. (For example, setting $h(x) = -\tau \|x\|_1$ for some parameter $\tau > 0$ tends to produce solutions x with few nonzeros.) An iteration scheme for this problem can be derived by replacing f with the function $\mathbf{A}_f$:

$$x^{t+1} := \arg\max_z \nabla f(x^t)^T (z - x^t) - \frac{1}{2\alpha_t} \|z - x^t\|_2^2 + h(z). \tag{15}$$

See [15] for application of this approach to compressed sensing problems, where f represents a linear-least squares loss. The paper [16] analyzes this approach for the case of f a logistic loss and h an ℓ_1 regularizer. This paper also describes acceleration of the

asymptotic convergence rate by using reduced Newton steps on the nonzero set identified by the basic step (15).

Methods that improve on first-order search directions are readily available for large-scale unconstrained optimization problems. L-BFGS ([21]), a limited-memory quasi-Newton method that requires storage of few vectors of length n, has already been tried with success ([23, 24]). Alternatives that could be tried include nonlinear conjugate-gradient and simple approaches such as heavy ball. (The latter method requires estimates of the extreme eigenvalues of the Hessian of f, but these may be readily available in some cases.)

We list here some possible directions for future research. Coordinate relaxation methods, in which steps are taken in just a subset of the variables on each iteration, have proved useful in other contexts and could be tried here. Another important class of optimization approaches that go by the general name of "stochastic gradient" methods [17] may also be useful for solving versions of these problems in which the data sets are large. Each iteration of these methods requires not an exact gradient of the loss function (which may entail a scan through the full data set) but rather an approximate gradient based on a small subset of the data. These approaches have proved highly effective in fast identification of approximate solutions to support-vector machine problems in machine learning; see for example [18].

The Rprop approach is amenable to some analysis [22] (and possibly improvement) using techniques from optimization. Around 1990, several optimization researchers analyzed back-propagation methods in terms of incremental gradient methods, and demonstrated an equivalence. The memory and path dependence that is inherent in the choice of steps in Rprop will make analysis more difficult.

8. CONCLUSIONS

In the paper we gave overview of some popular optimization methods for discriminative training in speech processing. Specifically we described EBW techniques and some associated with EBW methods. We also gave description of some general optimization methods that can be considered as alternatives to EM algorithms.

9. REFERENCES

[1] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "GIS-like Estimation of Log-Linear Models with Hidden Variables," in Proc. ICASSP, 2008.

[2] G. Heigold, S. Hahn, P. Lehnen, and H. Ney, "EM-Style Optimization of Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion," in Proc. ICASSP, 2011.

[3] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schlüter, "Equivalence of Generative and Log-Linear Models", IEEE Transactions on Audio, Speech & Language Processing, 2011, pp.1138-1148.

[4] G. Heigold, "A Log-Linear Discriminative Modeling Framework for Speech Recognition," Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, Jun. 2010.

[5] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo and A. Nadas, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems", IEEE Trans. Information Theory, 1991, vol 37(1).

[6] Y. Normandin, "An improved MMIE training algorithm for speakerindependent, small vocabulary, continuous speech recognition," in Proc. ICASSP, 1991.

[7] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, University of Cambridge, 2003.

[8] D. Kanevsky, D. Nahamoo, T. N. Sainath, B. Ramabhadran, and P. A. Olsen, "A-Functions: A Generalization of Extended Baum-Welch Transformations to Convex Optimization", in Proc. ICASSP, 2011.

[9] D. Kanevsky, T.N. Sainath, and B. Ramabhadran, "A generalized family of parameter estimation techniques", in Proc. ICASSP, 2009.

[10] D. Kanevsky, "Extended Baum transformations for general functions", in Proc. ICASSP, 2004

[11] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition", Speech Communication, vol. 34, pp.287-310, 2001.

[12] Liu, P. Liu, C. Jiang, H. Soong, F. Wang, R.-H., "A Constrained Line Search Optimization Method for Discriminative Training of HMMs", Audio, Speech, and Language Processing, IEEE Transactions, July 2008, Volume: 16 Issue: 5, pp. 900 - 909

[13] R. Hsiao and T. Schultz, "Generalized Baum-Welch Algorithm and its application to New Extended Baum-Welch Algorithm", in Proc. Interspeech, 2011.

[14] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, second edition, 1999.

[15] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," IEEE Transactions on Signal Processing, 57 (2009), pp. 2479-2493.

[16] S. J. Wright, "Accelerated block-coordinate relaxation for regularized optimization," Technical report, University of Wisconsin-Madison, August 2010. Revised September, 2011.

[17] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," SIAM J. Optim. 19 (2009), pp. 1574-1609.

[18] S. Shalev-Shwartz, Y. Singer and N. Srebro, "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM," In Proceedings of the 24th ICML, 2007.

[19] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The Rprop algorithm," In IEEE International Conference on Neural Networks (ICNN), 1993.

[20] E. McDermott et al., "Discriminative training for large vocabulary speech recognition using Minimum Classification Error," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 1, pp. 203 – 223, 2007.

[21] J. Nocedal and S. J. Wright. "Numerical Optimization" (2nd ed.), Berlin, New York: Springer-Verlag, ISBN 978-0-387-30303-1.

[22] A. D. Anastasiadis, G. D. Magoulas and M. N. Vrahatis, "New globally convergent training scheme based on the resilient propagation algorithm," Neurocomputing, vol. 64, pp. 253 – 270, 2005.

[23] J. Le Roux and E. McDermott, "Optimization methods for discriminative training," in Proc. Interspeech, 2005.

[24] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification," in Proc. Interspeech, 2005.