AUTOMATIC PARAMETER SETTING FOR ITERATIVE SHRINKAGE METHODS

*Raja Giryes, Michael Elad**

Technion - Israel Institute of Technology
Computer Science department
Haifa 32000, Israel
[raja,elad]@cs.technion.ac.il

Yonina C. Eldar

Technion - Israel Institute of Technology
Electrical Engineering department
Haifa 32000, Israel
yonina@ee.technion.ac.il

ABSTRACT

Linear inverse problems are very common in signal and image processing. Algorithms that solve such problems typically involve several unknown parameters that need to be tuned. Here we consider an iterated shrinkage method that is based on the separable surrogate functions (SSF) idea, which exploits the sparsity of the unknown vector in an appropriate representation. The key parameter controlling the algorithm's success is the prior weight, denoted λ. Previous work has addressed the automatic tuning of λ based on a generalized Stein Unbiased Risk Estimator (SURE) of the mean-squared error (MSE). The approach taken was to obtain a constant value of λ that leads to optimized results over a given set of iterations.

In this work we also rely on the generalized SURE, and propose an alternative, and highly effective method for tuning λ. Our algorithm chooses λ per iteration, based on the local estimated risk, considering the current iteration and a possible short look-ahead. We demonstrate this method and its superiority over the global approach both in terms of the resulting MSE and the convergence rate. We also show that the proposed scheme serves as a very reliable automatic halting mechanism for the iterative process.

***Index Terms*—** Iterated Shrinkage, Stein Unbiased Risk Estimator, Separable Surrogate Function, Inverse problem.

1. INTRODUCTION

In many applications in signal and image processing there is a need for solving a linear inverse problem. In such problems, an original deterministic signal $\mathbf{x}$ is deteriorated by a (known) linear operator $\mathbf{H}$, followed by additive white Gaussian noise $\mathbf{w}$ with variance σ^2. Given the measurement vector

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{1}$$

the goal is to reconstruct $\mathbf{x}$. We assume throughout that $\mathbf{H}$ has full column rank.

*This research was supported by the Goldstein UAV and Satellite Center.

There are many techniques in the literature for this task and most of them include parameters that require tuning. Tuning of the parameters is generally not an easy task. The typical objective in these problems is a recovery of a signal with minimal mean-squared error (MSE). However, in practice we do not have the original signal $\mathbf{x}$ to compare to, and thus we cannot know what choice of parameters minimizes the MSE. Therefore, the parameters are often tuned manually by looking at the reconstructed result.

The literature offers several automatic ways for choosing the parameters. One popular tool is the generalized cross validation (GCV) technique and its many variants [1]. An alternative method, shown to be superior to the GCV in many examples, is the Stein unbiased risk estimator (SURE) [2, 3]. SURE provides an unbiased estimate of the MSE for a candidate solver. Minimizing this estimated MSE, leads to the automatic tuning desired. However, the original SURE is limited to the case $\mathbf{H} = \mathbf{I}$, implying a denoising problem. One of the best known methods that uses SURE is Donoho's SureShrink denoising algorithm [4]. There are also other contributions along these lines, as in [5, 6, 7].

Recently, a generalization of SURE (GSURE) has been developed for more general models than (1) [8]. This provides a new opportunity for choosing the parameters automatically in a diverse set of inverse problems. A special case of these results has also been developed and used in [9] for the case of (1) with general $\mathbf{H}$ and white Gaussian noise.

In this paper we focus on automatic tuning for a specific reconstruction algorithm, following the work reported in [9]. The algorithm considered is an *iterated shrinkage* method, that is based on the separable surrogate functions (SSF) idea [10, 11, 12]. In this algorithm, the signal $\mathbf{x}$ is known to have a sparse representation over a dictionary $\mathbf{D}$. Reconstruction is obtained by an iterative estimator denoted $h_{\lambda,k}(\mathbf{y})$, where k is the iteration number, and both k and λ are to be tuned for best performance.

Previous work in [9] has already addressed the automatic tuning of λ based on GSURE. In their approach, for a fixed and pre-chosen number of iterations k_0, the GSURE is used to choose a fixed value of λ that optimizes the overall (estimated) MSE. For a fixed λ, the GSURE is also used again to determine the number of iterations. We refer to this method of setting λ and k as the *global method*.

In this work a different route it taken. We propose selecting the value of λ to minimize the estimated MSE of each iteration, thereby getting a different value of λ per iteration. We refer to this method as the *greedy method*. First, we show that this approach leads to faster convergence of the overall algorithm. Furthermore, another benefit of this strategy is a natural way for setting the number of iterations simultaneously – the iterations can be stopped when the MSE improvement is under a certain threshold. This method is also less sensitive to errors in the MSE estimator than the global approach. In order to further improve the performance with regard to the overall MSE in this greedy method, λ in each iteration can be chosen with a look-ahead on the MSE in the next few iterations.

This paper is organized as follows. In Section 2 the iterated shrinkage and SSF algorithms are explained. In Section 3 the generalized SURE is presented. Section 4 develops the use of GSURE for tuning λ and k. In Section 5 we present the benefits of the proposed greedy method and compare it to the global approach. Section 6 concludes the work and makes suggestions for further development.

2. THE SSF METHOD

In order to estimate the unknown signal, we assume that $\mathbf{x}$ has a sparse representation with respect to the dictionary $\mathbf{D}$. We further assume that the measured signal $\mathbf{y}$ is close to the reconstructed signal after applying the linear operator $\mathbf{H}$. Denoting by α the representation of the reconstructed signal, our reconstruction algorithm aims to minimize the objective function

$$\epsilon(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{HD}\alpha\right\|_2^2 + \lambda\left\|\alpha\right\|_1, \tag{2}$$

where λ, the relative weight of the two penalties in the objective, is unknown and needs to be adjusted.

In order to minimize the functional in (2), we use the SSF method [10, 11]. In this approach, a majorization of the objective function is obtained by adding the term

$$\frac{c}{2}\left\|\alpha - \alpha_0\right\|_2^2 + \frac{1}{2}\left\|\mathbf{HD}\alpha - \mathbf{HD}\alpha_0\right\|_2^2,$$

where α_0 is an arbitrary vector and c must satisfy $c > \left\|(\mathbf{HD})^T\mathbf{HD}\right\|_2$. The solution of the objective after the addition is given in closed-form by

$$\alpha_{opt} = S_{1,\lambda/c}\left(\frac{1}{c}(\mathbf{HD})^T(\mathbf{y} - \mathbf{HD}\alpha_0) + \alpha_0\right). \tag{3}$$

The operator $S_{1,\lambda/c}$ is an element-wise soft thresholding function with threshold value λ/c. Throughout the paper, we use a smooth approximation of the soft-threshold, as proposed in Section 2.2 in [11]. The sequence

$$\alpha_{i+1} = S_{1,\lambda/c}\left(\frac{1}{c}(\mathbf{HD})^T(\mathbf{y} - \mathbf{HD}\alpha_i) + \alpha_i\right), \tag{4}$$

is proven to converge to the minimizer of (2). This leads to the estimator

$$\hat{\mathbf{x}} = h_{\lambda,k}(\mathbf{y}) = \mathbf{D}\alpha_k. \tag{5}$$

In practice, the parameter λ and the number of iterations k need adjustment for better estimation of the original signal $\mathbf{x}$. We now turn to present the generalized-SURE, which will be used for this task.

3. ESTIMATING THE MSE – GSURE

The most popular criterion for evaluating performance of any estimator is the MSE. Given an estimator $\hat{\mathbf{x}}$ of a signal $\mathbf{x}$ the MSE is given by

$$MSE(\hat{\mathbf{x}}) = E\left[\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|_2^2\right], \tag{6}$$

where the estimator is a function of the measurement $\mathbf{y}$ in (1). It is well known that for every estimator (being a function of the measurements), there is an alternative estimator with equal or smaller MSE, which is a function of the sufficient statistics (Rao-Blackwell theorem [13]). Thus we will look at $\hat{\mathbf{x}} = h(\mathbf{u})$ where $\mathbf{u} = (1/\sigma^2)\mathbf{H}^T\mathbf{y}$ is the sufficient statistic for the model (1).

The MSE of an estimator $h(\mathbf{u})$ can be written as follows:

$$\begin{aligned} & E\left[\left\|\mathbf{x} - h(\mathbf{u})\right\|_2^2\right] = \\ & = \left\|\mathbf{x}\right\|_2^2 - 2E\left[\mathbf{x}^T h(\mathbf{u})\right] + E\left[\left\|h(\mathbf{u})\right\|_2^2\right]. \end{aligned} \tag{7}$$

Since the MSE depends on the unknown $\mathbf{x}$, we would like to replace it by an unbiased estimate $e(h(\mathbf{u}))$, that is $E[e(h(\mathbf{u}))] = E[\|\mathbf{x} - h(\mathbf{u})\|_2^2]$. A result developed in [8] shows that

$$E[\mathbf{x}^T h(\mathbf{u}))] = -E[div_{\mathbf{u}}(h(\mathbf{u})) - \mathbf{x}_{ML}^T h(\mathbf{u})], \tag{8}$$

where div is the divergence operator:

$$div_{\mathbf{u}}(h(\mathbf{u})) = \sum_{i=1}^{n}\frac{dh_i(\mathbf{u})}{du_i}, \tag{9}$$

and $\mathbf{x}_{ML}$ denotes the maximum likelihood estimator for the problem (1):

$$\mathbf{x}_{ML} = \mathbf{H}^{\dagger}\mathbf{y} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{u}, \tag{10}$$

where $\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is the Moore-Penrose pseudo inverse. Using (8) leads to the generalized SURE (GSURE) proposed in [8],

$$\begin{aligned} e(h(\mathbf{u})) = & \left\|\mathbf{x}\right\|_2^2 + \left\|h(\mathbf{u})\right\|_2^2 \\ & + 2div_{\mathbf{u}}(h(\mathbf{u})) - 2\mathbf{x}_{ML}^T h(\mathbf{u}). \end{aligned} \tag{11}$$

Since $\mathbf{x}$ is a constant, independent of the choice of estimate, it will not effect the minimization of $e(h(\mathbf{u}))$.

4. PARAMETER TUNING ALGORITHMS

In order to use GSURE with the iterated shrinkage algorithm, described in Section 2, we need to reformulate the iterative equation as a function of $\mathbf{u}$ instead of $\mathbf{y}$. Rewriting (4) leads to

$$\alpha_{i+1} = S_{1,\lambda/c}\left(\frac{1}{c}\mathbf{D}^T(\sigma^2\mathbf{u} - \mathbf{H}^T\mathbf{H}\mathbf{D}\alpha_i) + \alpha_i\right), \quad (12)$$

and $\hat{\mathbf{x}} = h_{\lambda,k}(\mathbf{u}) = \mathbf{D}\alpha_k$.

In order to tune the algorithm parameters, we estimate the MSE using (11). The derivatives of $\hat{\mathbf{x}}$ are calculated recursively by first noticing that

$$\frac{d\alpha_k}{d\mathbf{u}} = S'_{1,\lambda/c}\left(\frac{1}{c}\mathbf{D}^T(\mathbf{u} - \mathbf{H}^T\mathbf{H}\mathbf{D}\alpha_{k-1}) + \alpha_{k-1}\right) \cdot \quad (13)$$
$$\left[\frac{\sigma^2}{c}\mathbf{D}^T - \frac{1}{c}(\mathbf{H}\mathbf{D})^T\mathbf{H}\mathbf{D}\frac{d\alpha_{k-1}}{d\mathbf{u}} + \frac{d\alpha_{k-1}}{d\mathbf{u}}\right],$$

where $S'_{1,\lambda/c}(\cdot)$ is an element wise derivative of the thresholding function, organized as a diagonal matrix. From here, the divergence of the estimator can be directly obtained by multiplying the above by $\mathbf{D}$ from the left, gathering the diagonal of the matrix and summing it up. Now that we hold a complete expression for the GSURE MSE estimate, k and λ can be chosen by minimizing this expression.

The first method we describe is the global method, developed in [9]. For a fixed and pre-chosen number of iterations k_0, the λ which minimizes the GSURE expression is chosen. Repeating this process for various values of k_0, one can minimize the overall global MSE with respect to both k_0 and λ. As an analytical minimization of the GSURE is hard to achieve, we use a golden section search [14].

4.1. The Greedy Method

Turning to the local alternative, instead of using a constant λ for all the iterations, the value of λ can be determined as the one that minimizes the estimated MSE of the current iteration. The algorithm proposed is the following:

- Initialize α_0 and calculate its derivative w.r.t. $\mathbf{u}$.
- Repeat:
  1. Set $\lambda_i^* = \arg\min_\lambda(e(h_\lambda(\mathbf{u}, \alpha_{i-1})))$.
  2. Perform the iteration in (12) for the calculation of α_i using λ_i^*.
  3. Compute $MSE_i^* = e(h_{\lambda_i^*}(\mathbf{u}, \alpha_{i-1}))$.
  4. If $MSE_{i-1}^* - MSE_i^* \leq \delta$, stop.
- $\hat{\mathbf{x}} = \mathbf{D}\alpha_k$.

The complexity of the greedy and the global methods are the same. Denoting by T the GSURE calculation time (per-iteration), n_{gs} the golden-section number of iterations, and n_{is} the number of iterations of the iterated shrinkage algorithm, the time complexity of both methods is $O(n_{is}n_{gs}T)$. The difference is that in the greedy method the number of iterations is set automatically during the setting of λ_i, whereas in the global method the number of iterations is either pre-chosen and thus suboptimal, or searched using yet another golden-search for optimizing k, increasing the overall complexity.

4.2. Look-Ahead Greedy Method

Simulations demonstrate that the greedy method achieves an overall MSE close to the one achieved by the global method, but with faster convergence. To further decrease the MSE, we can modify the greedy method by introducing a look ahead. One of the problems of the greedy method is that it minimizes the MSE of the current iteration but can harm the overall MSE. Thus, instead of choosing λ that minimizes the estimated MSE of the current iteration, it can be chosen as the one that minimizes the estimated MSE of r iterations ahead, assuming that these r iterations are performed using the greedy approach described above. This change provides a look-ahead of r iterations, formulated as the following algorithm:

- Initialize α_0 and calculate its derivative w.r.t. $\mathbf{u}$.
- Repeat:
  1. Set λ_i^* by minimizing the estimated MSE r iterations ahead using the above-described greedy algorithm.
  2. Perform a single iteration as in (12) for the calculation of α_i using λ_i^*.
  3. Compute $MSE_i^* = e(h_{\lambda_i^*}(\mathbf{u}, \alpha_{i-1}))$.
  4. If $MSE_{i-1}^* - MSE_i^* \leq \delta$, stop.
- $\hat{\mathbf{x}} = \mathbf{D}\alpha_k$.

In step 1, for each λ checked in the golden-section for the current iteration, r iterations of the greedy method are performed, as described above. Finally, λ for the current iteration is chosen such that the estimated MSE of the last rth greedy iteration is minimized.

The time complexity of the r look-ahead greedy method is $n_{is}(n_{gs})^2rT$, which is theoretically $n_{gs}r$ times slower than the other two methods. However, n_{is} of the look-ahead method is expected to be smaller due to fast convergence, while also achieving lower MSE, as we illustrate next.

5. RESULTS

In order to compare the methods discussed, we performed a synthetic experiment, with a 1D piecewise constant signal of length $n = 256$ samples. This signal contains 13 steps in random locations and random amplitudes, and its power per entry[1] is 2.55. The undecimated Haar transform is chosen as the dictionary, and the operator $\mathbf{H}$ is chosen as the square matrix representing a blur operator using the filter $[1, 2, 4, 2, 1]$ (ℓ_1 normalized). The variance of the white Gaussian noise is $\sigma = \sqrt{2}$.

The first test verifies that GSURE provides a good estimate of the MSE when using the greedy method. In Figure 1 a comparison between choosing the λ value in the greedy method based on the GSURE and based on the true MSE is shown. In each of the two greedy methods, we estimate the original signal. The graphs in the figure present the MSE achieved in each such iteration of the two versions. We see that the MSE obtained with GSURE is very close to the one achieved using the true MSE, demonstrating the reliability of GSURE.

In Figure 2 the MSE after each iteration of the greedy method (when λ is chosen based on GSURE) is presented together with its GSURE estimation. The estimation has the same behavior as the true MSE. It can be seen that in both graphes the minimum (marked in a circle) are very close. Choosing the number of iterations based on GSURE yields an MSE very close to the one based on the real MSE. This justifies our claim about using the greedy method as an automatic stopping rule, by detecting an increase in the MSE, or a small improvement.

Comparison of the MSE of the global, the greedy, and the look-ahead methods is presented in Figure 3. For the global method, for each number of iterations, a different λ value that minimizes the estimation of the MSE is being chosen. It can be observed that the greedy method converges faster than the global method and with a relatively small loss (for a large number of iterations). Note that when the number of iterations is smaller than 50 in this test, the greedy method is actually better.

The look-ahead greedy method with $r = 1$ and $r = 2$ achieves a lower MSE, compared to the previous two alternatives, while also enjoying fast convergence. As expected, longer look-ahead results in smaller MSE, at the price of higher complexity.

Another benefit of the greedy methods observed during our simulations is numerical stability. Because the choice of the parameters in the global method relies on one GSURE estimation only, the global method is more sensitive to errors in the estimation of the MSE. In contrast, in the greedy methods, in each iteration a new estimation of the MSE is obtained and thus estimation error in one iteration has only a local effect that can be recovered in the next iterations.

[1]This is defined as $\sqrt{\|\mathbf{x}\|_2^2/n}$.

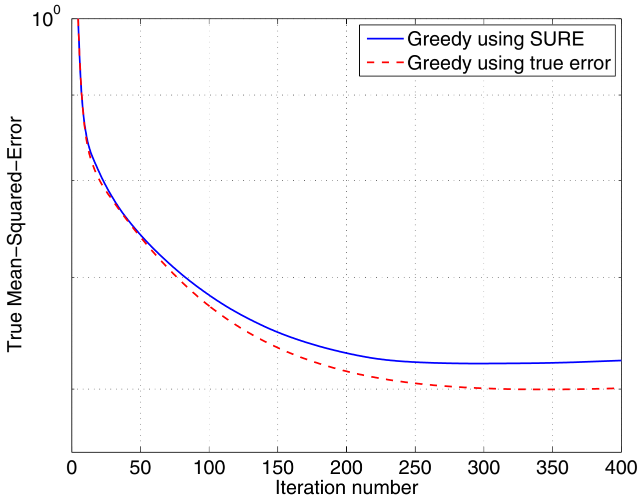

Fig. 1. The true MSE as a function of the iteration, for two versions of the greedy algorithm: (i) choosing λ based on the true MSE; and (ii) choosing λ based on GSURE.

6. CONCLUSION AND DISCUSSION

This paper considers an automatic tuning of parameters for inverse problems. An automatic way for choosing λ in each iteration in the iterated shrinkage algorithm is considered and demonstrated. We compare the global tuning developed in [9], to two greedy methods - a simple and a look-ahead version. The two greedy methods are shown to outperform the global approach in the final MSE obtained, the speed of convergence, and the ability to set an automatic stopping rule.

The look-ahead method can be extended in various ways. A variation of it can be used in a coordinate-descent form, optimizing the λ value for each iteration while optimizing the global estimated MSE. Other variations can be considered based on the idea of choosing different λ in each iteration and basing the choice on different stages of the algorithm. The methods proposed here were checked on 1D signals of low-dimensions; their application to high-dimensional signals and images is being considered as a follow-up work.

7. REFERENCES

[1] G.H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, May 1979.

[2] C.M. Stein, "Estimation of the mean of a multivariate distribution," *Proc. Prague Symp. Asymptotic Statist.*, pp. 345–381, 1973.

[3] C.M. Stein, "Estimation of the mean of a multivariate

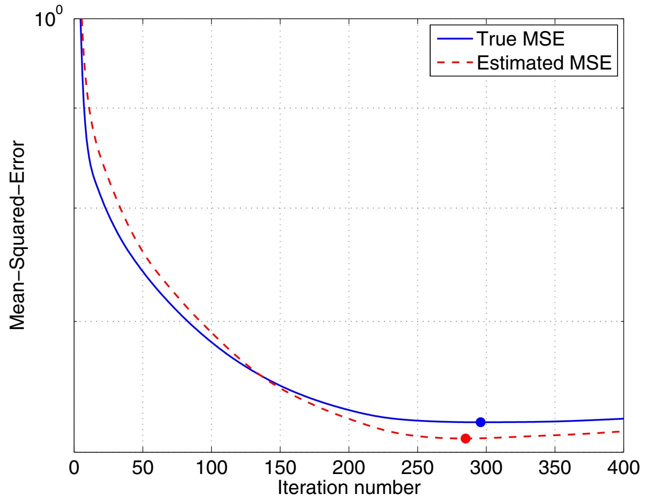

Fig. 2. The GSURE-estimated and the true MSE as a function of the iteration number for the greedy algorithm. The circles denote the minimum point for the two graphs.

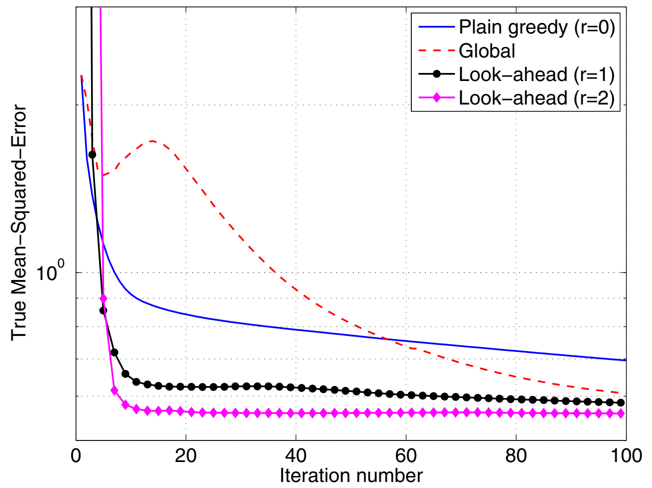

Fig. 3. The true MSE as a function of the iteration number for the global, greedy, and look-ahead-greedy methods.

normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[4] D.L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200-1224, Dec. 1995.

[5] T. Blu, F. Luisier, "The SURE-LET approach to image denoising," *IEEE Trans. on Image Process.*, vol. 16, no. 11, pp. 2778–2786, Nov. 2007.

[6] J.-C. Pesquet and D. Leporini, "A new wavelet estimator for image denoising," in *Proc. 6th Int. Conf. Image Processing and Its Applications*, Jul. 14-17, 1997, vol. 1, pp. 249-253.

[7] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Steins principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814-1830, Nov. 2005.

[8] Y.C. Eldar, "Generalized SURE for exponential families: Applications to regularization," to appear in *IEEE Trans. on Signal Processing*.

[9] C. Vonesch, S. Ramani and M. Unser, "Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint", to appear in *IEEE ICIP*, 2008.

[10] I. Daubechies, M. Defrise, and C. De-Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, LVII:1413–1457, 2004.

[11] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, pp. 346–367, Nov. 2007.

[12] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky, "A Wide-Angle View at Iterated Shrinkage Algorithms", *SPIE (Wavelet XII)*, San-Diego CA, August 26-29, 2007.

[13] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Upper Saddle River, NJ: Prentice Hall, Inc., 1993.

[14] J. Kiefer, "Sequential minimax search for a maximum," *Proceedings of the American Mathematical Society*, vol. 4, pp. 502–506, 1953.