A Cross-Entropy Method that Optimizes Partially Decomposable Problems: A New Way to Interpret NMR Spectra

Siamak (Mohsen) Ravanbakhsh, Barnabás Póczos and **Russell Greiner**
Computing Science Department, University of Alberta
Edmonton, Alberta, Canada
{mravanba, poczos, rgreiner}@ualberta.ca

Abstract

Some real-world problems are *partially decomposable*, in that they can be decomposed into a set of coupled sub-problems, that are each relatively easy to solve. However, when these sub-problem share some common variables, it is not sufficient to simply solve each sub-problem in isolation. We develop a technology for such problems, and use it to address the challenge of finding the concentrations of the chemicals that appear in a complex mixture, based on its one-dimensional ^{1}H Nuclear Magnetic Resonance (NMR) spectrum. As each chemical involves clusters of spatially localized peaks, this requires finding the shifts for the clusters and the concentrations of the chemicals, that collectively produce the best match to the observed NMR spectrum. Here, each sub-problem requires finding the chemical concentrations and cluster shifts that can appear within a limited spectrum range; these are coupled as these limited regions can share many chemicals, and so must agree on the concentrations and cluster shifts of the common chemicals. This task motivates CEED: a novel extension to the Cross-Entropy stochastic optimization method constructed to address such partially decomposable problems. Our experimental results in the NMR task show that our CEED system is superior to other well-known optimization methods, and indeed produces the best-known results in this important, real-world application.

1. Introduction

Our practical goal is to automatically "interpret" a ^{1}H Nuclear Magnetic Resonance (NMR) spectrum, based on a library of the "signatures" of a set of chemicals — that is, to find the concentrations of the chemicals (think "linear combination of these signatures") that best matches the spectrum. The challenge is that each signature is actually a set of clusters, where each cluster (a set of peaks with known relative positions and heights) can shift within a small region. Hence, we can view this task as a multi-extremal continuous optimization problem that can involves hundreds of bounded variables (corresponding to the concentrations of the chemicals, and the shifts of the clusters). As the spectra is noisy and the loss function is not convex, the best candidates for solving this problem seem to be global, stochastic optimization methods. One such technique is the *Cross Entropy* (CE) method — a relatively new method that has been successfully applied to many different domains, including a variety of continuous and combinatorial optimizations such as clustering and vector quantization, policy optimization, and buffer allocation (see (Rubinstein and Kroese 2004) and references therein). While the standard CE has proven sufficient for many of these tasks, in some situations we may be able to get yet better performance by exploiting some characteristic of the problem itself. Motivated by our real-world NMR task, we develop a variation of the original CE method that can exploit the structure of partially decomposable problems, by an iterative process that, in each step, first finds a distribution of "solutions" to each sub-problem separately and then combines these distributions over each variable domain. While there are many primal and dual decomposition methods in the convex optimization literature that also try to solve several sub-problems simultaneously (Boyd and Vandenberghe 2004), we are addressing a non-convex problem using a non-deterministic method. Similarly, genetic algorithms have been used to improve probabilistic models that exploit inter-variable dependencies to control future cross-overs or samplings (see for example (Baluja 2002)). In addition to methodological differences, our work differs by considering interdependencies of (sub)problems, rather than just relations of variables.

Section 2 quickly summarizes the basic CE method, then uses the SAT problem to motivate our extension, called *CE Exploiting partial Decomposability* (CEED). Section 3 provides experimental results showing that CEED works effectively on the challenging and important problem of interpreting NMR spectra. It also compares CEED's performance to other optimization methods, including gradient descent, simulated annealing and genetic algorithms, and a current state-of-the-art system for NMR analysis. The web-page (RPG 2010) provides some more background on CE, the results of applying CEED to SAT and Sudoku problems, and additional results on the analysis of NMR spectra.

2. Extending CE to Exploit Decomposability

2.1 Introduction to the CE Method

We are given a loss function $L : \mathcal{X} \to \Re$ defined over a bounded convex (continuous or discrete) domain $\mathcal{X}$. Our

goal is to find an approximation to the global minimizer of L: $x^* \doteq \arg\min_{x\in\mathcal{X}} L(x)$. The Ordinary Cross Entropy (OCE) method addresses this using an iterative process that involves sampling from distributions that have the least Cross Entropy (KL-divergence) to distributions over successively improving level sets.

Let the indicator function $\mathbb{I}_{\{\text{condition}\}}$ be 1 if the condition is true and 0 otherwise. Define a uniform distribution over level-set $L(x) < \tau$: $h_\tau(x) \doteq \frac{\mathbb{I}_{\{L(x)<\tau\}}}{\int_{\mathcal{X}} \mathbb{I}_{\{L(x_1)<\tau\}} dx_1}$. Let $f(\cdot)$ be the distribution that minimizes the cross-entropy to h_τ:

$$\begin{aligned} f(\cdot) &\doteq \arg\min_g \left\{ \int_{\mathcal{X}} h_\tau(x) \ln \frac{h_\tau(x)}{g(x)} dx \right\} \\ &= \arg\max_g \left\{ \int_{\mathcal{X}} h_\tau(x) \ln(g(x))\, dx \right\} \end{aligned} \quad (1)$$

The OCE method seeks f for a small τ, in which case $\hat{x}^* = mode(f)$ is typically a good approximation to x^*.

This is performed iteratively, by adaptively updating τ and f; we use τ^t and f^t to denote their values on iteration t. In iteration t, OCE first draws N instances from f^{t-1}, $\mathbf{X} = \{X_n \sim f^{t-1}(x)\}_{1\le n\le N}$, renumbered such that $L(X_1) \le \ldots \le L(X_N)$. It then uses the *elite samples* $\mathbf{X}_{elite} = \{X_n\}_{1\le n\le \lceil \rho N\rceil}$, which are the top ρ fraction of the instances. At the next time-step, OCE sets $\tau^t = L(X_{\lceil \rho N\rceil})$ to be the smallest level-set that includes these top ρ fraction of the instances, and then use $\mathbf{X}_{elite}$ as an empirical approximation for $h_{\tau^t}(x)$. OCE then uses the empirical version of Eq 1 to calculate f^t.

We use a parametric approach by restricting f to a parametric family $\mathcal{F} \doteq \{f_v(x)\}_v$. In this parametric representation, the empirical counterpart of Eq 1 at time-step t simplifies to

$$\begin{aligned} \hat{v}^t &\doteq \arg\max_v \left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\{L(X_n)<\tau^t\}} \ln f_v(X_n) \right\} \\ &= \arg\max_v \left\{ \sum_{X_n \in \mathbf{X}_{elite}} \ln f_v(X_n) \right\}, \end{aligned} \quad (2)$$

where $X_n \sim f_{v^{t-1}(x)}$ are the instances generated from distribution that has the minimum cross-entropy to the previous level set τ^{t-1}. The solution to Eq 2 is the *maximum likelihood* estimate of parameter v^t from the elite samples. This update equation has an analytical form for the natural exponential family, which makes this method practical (Rubinstein and Kroese 2004).

2.2 Decomposability

Many real-world optimization problems can be (partially) decomposed into a set of 'simpler' sub-problems, each with fewer variables. As most problem domains grow exponentially in the number of variables, this reduction in the number of variables can be very advantageous. As a trivial example, consider SAT: Given a Conjunctive Normal Form formula (over D variables with P clauses), find an assignment that satisfies every clause (Garey and Johnson 1990). Here, we can view each individual clause $L_p = \pm_{d_1} X_{d_1} \vee \cdots \vee \pm_{d_k} X_{d_k}$ as a sub-problem, and seek an assignment that satisfies it alone. Of course, this is trivial — indeed, a $1 - 1/2^k$ portion of assignments satisfy each individual L_p. We could do this for each clause individually, to obtain several assignments over (various subsets of) the variables. The challenge then is combining these resulting assignments, to find a single assignment that satisfies every clause. Our novel CE-based approach, CEED, deals with several *distributions* over each variable (here one associated with each clause that includes that variable), and provides a way to combine them to produce a single good distribution for each variable.

We assume there are D variables, each parameterized by Q parameters. (*E.g.*, in the SAT problem, each of the D variables is represented by a Bernoulli variable v_d, with two parameters $Q = 2$ representing the probability of that variable being true or false.) More formally, let $\mathcal{X}^D$ be a bounded D-dimensional domain representing the possible tuples of variable assignments, and for each $d \in \{1, ..., D\}$, let

$$v_d = [v_d(1), ..., v_d(Q)] \quad \in \quad \Re^Q \quad (3)$$

be a row vector representing the parameters for that variable.

Let $\mathcal{F} = \{f_v\}$ be a joint distribution of a specific family defined over $\mathcal{X}^D$: $f_v(x) \doteq \prod_{d=1}^{D} f_{v_d}(x_d)$, where for each $d \in \{1, ..., D\}$, v_d is the parameter vector of the distribution $f_{v_d} : \mathcal{X} \to \Re$, and $v = [v_1; ...; v_D] \in \Re^{DQ}$ is the parameter vector of $f_v : \mathcal{X}^D \to \Re$. For combinatorial optimizations, that is when $\mathcal{X} = \{1, ..., Q\}$ is discrete, $f_{v_d}(x_d = q) = \Pr(x_d = q) = v_d(q)$ is a probability mass function (pmf), where v_d is located on a simplex — *i.e.*, $\sum_{q=1}^{|\mathcal{X}|} v_d(q) = 1$ and $v_d(q) \ge 0$.

For notation, we let $x_{\mathcal{A}}$ (resp., $v_{\mathcal{A}}$) be the restriction of x (resp., v) to the coordinates in $\mathcal{A} \subseteq \{1, ..., D\}$. Using the same notation, $f_{v_{\mathcal{A}}}(x_{\mathcal{A}}) = \prod_{d\in\mathcal{A}} f_{v_d}(x_d)$ is a joint distribution defined over $\mathcal{X}^D_{\mathcal{A}}$, the restriction of $\mathcal{X}^D$ to $\mathcal{A}$.

For each $p \in \{1, ..., P\}$, let $\mathcal{M}(p) \subseteq \{1, ..., D\}$ index the set of variables in the p^{th} group (*e.g.*, this could be the indices of the Boolean variables in the p^{th} clause). Observe that, for any loss function of the form $L(x) = \sum_{p=1}^{P} L_p(x_{\mathcal{M}(p)})$ and for any value τ, if $\sum_p \tau_p \le \tau$ then we have $\mathbb{I}_{\{L(x)<\tau\}} \ge \prod_{p=1}^{P} \mathbb{I}_{\{L(x_{\mathcal{M}(p)})<\tau_p\}}$, for some values of $\{\tau_p\}$. We can therefore try to solve the optimization Eq 2 by finding assignments that produce good results for all sub-problems at the same time.

Given this notion of sub-problem, we can show the interaction of variables in sub-problems using the *coupling matrix* $C \in \{0, 1\}^{P\times D}$, whose rows correspond to sub-problems and columns to variables, where the element $C_{p,d}$ is 1 iff the sub-problem p depends on variable x_d. We then define $\mathcal{M}(p) = \{d : C_{p,d} = 1\}$ to index the *variables* involved in *sub-problem* p and $\mathcal{A}(d) = \{p : C_{p,d} = 1\}$ to index *sub-problems* that involve *variable* x_d.[1]

[1]To summarize the notation: We have D variables, grouped into P (overlapping) subgroups. $f^t_{v^t_{d,p}}(\cdot)$ refers to a distribution where the superscript "^{t}" indexes the iteration (here for v and f, and also for the level $\tau \in \Re$), v stands for the parameters, where

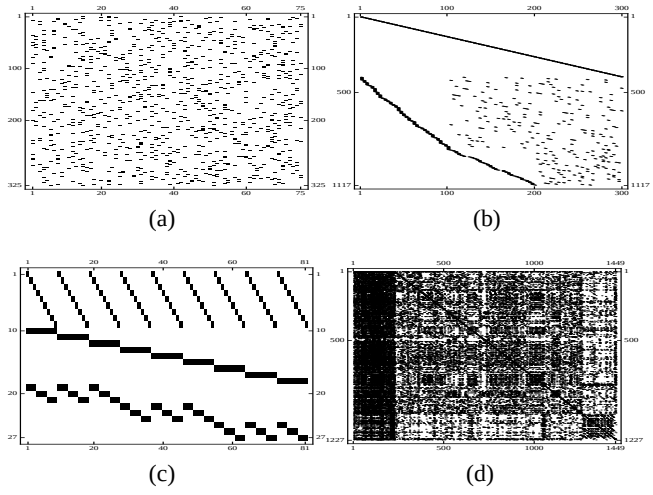

Figure 1: Coupling matrices for (a) Random SAT, (b) Structured SAT, (c) Sudoku, (d) NMR problems

Figures 1(a) and 1(b) show the coupling matrix for a random SAT and a SAT formulated graph coloring problem; here each clause is a sub-problem. Figure 1(c) shows the coupling matrix for the formulation of a 9×9 Sudoku problem, where the sub-function to be maximized (in each sub-problem) is the number of *distinct* elements of each row, column and sub-square of the Sudoku puzzle . (Each element of the puzzle is represented by a pmf over $\{1, ..., 9\}$.) This matrix clearly shows how the various sub-problems are inter-related. Finally, Figure 1(d) shows the coupling matrix for our NMR Spectra interpretation task; see Section 3.

2.3 Cross Entropy Exploiting partial Decomposability (CEED) method

The basic idea underlying the CEED method is to first obtain several estimates of each variable using the loss functions for each sub-problem, and then combine these estimates into a single distribution, which CEED uses for sampling in the next iteration. Note we identify each sub-problem with its associated loss function L_p; also the number of instances drawn for each sub-problem N_p is computed based on the complexity of that sub-problem. (This complexity varies by application; see below.)

We can summarize CEED algorithm in the following steps:

Input: Prior $\hat{v}^0$; Elite fraction $\rho \in (0, 1)$;
Set of sub-problem/#instances pairs $\{\langle L_p, N_p \rangle\}$
Output: approximate minimizer of $\sum_p L_p(x)$, $\hat{x}^* \in \mathcal{X}^D$

0. Initialize $t = 1$, and derive $\mathcal{A}(d)$ and $\mathcal{M}(p)$ using the domain of given sub-problem $L_p(.)$

1. At time-step t, draw N_p instances from the joint distribution $f_{\hat{v}^{t-1}_{\mathcal{M}(p)}}$, for each sub-problem L_p.

2. For each sub-problem p, calculate the error for the instances, and find a joint distribution with the least cross entropy to the elite instances (the top ρ fraction). More formally, similar to Eq 2, define, for each $p \in \{1, ..., P\}$,

$$\hat{v}^t_{\mathcal{M}(p),p} = \arg\max_{v} \left\{ \frac{1}{N_p} \sum_{n=1}^{N_p} \mathbb{I}_p(X_n) \ln f_{v_{\mathcal{M}(p)}}(X_n) \right\} \quad (4)$$

where each $X_n \sim f_{\hat{v}^{t-1}_{\mathcal{M}(p)}}$. Here, $\hat{v}^t_{\mathcal{M}(p),p}$ is the estimate of many distributions, each parameterized by $v^*_{\mathcal{M}(p)}$ (the restriction of global optima to sub-problem p) for all of the variables in $\mathcal{M}(p)$, given by sub-problem p at time-step t.

3. Each variable v_d appears in each of the sub-problems in $\mathcal{A}(d)$, and so is estimated by multiple parameter values $\{\hat{v}^t_{d,p} \mid p \in \mathcal{A}(d)\}$ — one for each sub-problem $p \in \mathcal{A}(d)$. We therefore "combine" the associated distributions $\{f_{\hat{v}^t_{d,p}} \mid p \in \mathcal{A}(d)\}$ to produce a single parameter value $\hat{v}^t_d$, then join this over all variables v_d to produce a single parameter value $\hat{v}^t$, describing a single joint distribution $f_{\hat{v}^t}$ over all variables. See Section 2.4.

4. If not converged, increment t and return to Step 1. Else stop and return the mode of $f_{\hat{v}^t}$ as an approximation to x^*.

$v_{d,p}$ refers to the parameters of a distribution learned for the d^{th} variable, based on the p^{th} sub-problem. Note the first subscript (called "d" above) can be a set — *e.g.,* $v_{A,p}$ refers to the parameters learned for the set of variable indices A, which could be $A = \mathcal{M}(p) \subset \{1, ..., D\}$.

For our SAT example, after sampling a distribution, and assigning a distribution to the variables in each clause, we obtain several distributions over each variable X — one joint distribution associated with each clause that contains X. We then combine these distributions to produce a single distribution over each variable, which we then use in the next iteration.

In terms of the general coupling matrix C, each CEED iteration first estimates the parameters in each row $\{\hat{v}^t_{\mathcal{M}(p),p}\}_{1 \le p \le P}$, then combines the estimates in each column, $\{\{\hat{v}^t_{d,p}\}_{p \in \mathcal{A}(d)}\}_{1 \le d \le D}$ to get $\hat{v}_t$. This is repeated in each iteration, using the previous estimate as the sampling distribution.

Step 3 needs to combine several estimates. We know that each estimate $\hat{v}_{d,p}$ of variable d given by sub-problem p corresponds to a distribution, $f_{\hat{v}_{d,p}}(x)$. Here, CEED uses the distribution that minimizes the sum of the KL-divergences to the given set of distributions:

$$\begin{aligned} \hat{v}^t_d &= \arg\min_{\omega} \left\{ \sum_{p \in \mathcal{A}(d)} \mathcal{D}_{KL}(f_\omega, f_{\hat{v}^t_{d,p}}) \right\} \quad (5) \\ &= \arg\min_{\omega} \left\{ \sum_{p \in \mathcal{A}(d)} \left(\int_{x \in \mathcal{X}} f_\omega(x) \ln(\frac{f_\omega(x)}{f_{\hat{v}^t_{d,p}}(x)}) dx \right) \right\} \end{aligned}$$

Although this optimization (Eq 5) is convex for the exponential family, it is still challenging to find the global optima

here. The next section therefore provides a linear combination method to approximate this solution to Eq 5. We believe this is a novel method for combining experts' votes when each expert is reporting a maximum likelihood estimate.[2]

2.4 Combining ML Estimates Using Their Fisher Information

Given a set of random *i.i.d.* instances $\mathbf{X} = \{X^1, ..., X^N\}$, the *score function* is defined as the gradient of the log-likelihood function: $U(v, \mathbf{X}) = \frac{\partial \log(\mathcal{L}(v;\mathbf{X}))}{\partial v}$, where $\mathcal{L}(v;\mathbf{X}) = \prod_{n=1}^{N} f_v(X^n)$ is the likelihood function based on the parameter v. This is the zero vector for the maximum-likelihood (ML) parameters. Since for ML estimates $\mathbb{E}\{U(v, X)|v\} = 0$, the variance of the score function is the quantity that contains the information about the ML estimate — *a.k.a.* its *Fisher information*:

$$\mathcal{I}(v) \;=\; \mathbb{E}\{U(v,X)^2|v\} \;=\; \mathbb{E}\left\{[\frac{\partial}{\partial v}\log(f_v(X))]^2 \mid v\right\}$$

If the distribution family $\mathcal{F}$ satisfies the regularity condition $\int \frac{\partial^2}{\partial v^2} f_v(x)\, dx \;=\; 0$, then we can also write the Fisher information as (Lehmann and Casella 1998):

$$\mathcal{I}(v) \;=\; -\mathbb{E}\left\{\frac{\partial}{\partial v}U(v,X) \mid v\right\} \;=\; -\mathbb{E}\left\{\frac{\partial^2}{\partial v^2}\log(f_v(X)) \mid v\right\}$$

The element $\mathcal{I}_{d,d'}(\tilde{v})$ (*i.e.*, the element in row d and column d' of the matrix $\mathcal{I}$) basically shows the rate of change in the maximum likelihood estimation of parameter $v(d) \in \Re$ by changing the parameter $v(d')$ in the neighborhood of $\tilde{v}(d)$; see Eq 3. Therefore higher values suggest greater accuracy in the maximum likelihood estimation. This is the basis of our approach for combining estimations; we combine the estimations $\{\hat{v}^t_{d,p}\}_p$ by weighting the estimates with the Fisher information matrices $\mathcal{I}(\hat{v}^t_{d,p})$:

$$\hat{v}^t_d \;\doteq\; \left(\sum_{p\in\mathcal{A}(d)} \hat{v}^t_{d,p}\mathcal{I}(\hat{v}^t_{d,p})\right)\left(\sum_{p\in\mathcal{A}(d)} \mathcal{I}(\hat{v}^t_{d,p})\right)^{-1} \qquad (6)$$

We can also connect Eq 6 with the minimization Eq 5 by using the relation (Kullback 1959): $2\mathcal{D}_{KL}(f_v, f_{v+dv}) \approx dv^T\, \mathcal{I}(v)\, dv$. This provides a second degree approximation of KL-divergence, which we can use to approximate the solution to the optimization (Eq 5):

$$\begin{aligned}\hat{v}^t_d \;&=\; \arg\min_{\omega}\left\{\sum_{p\in\mathcal{A}(d)} \mathcal{D}_{KL}(f_\omega, f_{\hat{v}^t_{d,p}})\right\}\\ &\approx\; \arg\min_{\omega}\left\{\sum_{p\in\mathcal{A}(d)} \frac{1}{2}(\hat{v}^t_{d,p}-\omega)^T\mathcal{I}(\hat{v}^t_{d,p})(\hat{v}^t_{d,p}-\omega)\right\}\end{aligned}$$

This minimization has the analytical solution, in the form of the linear combination given by Eq 6.

[2]Fisher (1925) was the first to use *sample weighting* proportional to the Fisher information.

We can also relate combination of ML estimates using their Fisher Information to the well-known method of linear combination by the inverse of the variance-covariance matrix. The *Cramér-Rao* theorem (Cramer 1946; Rao 1945) states that any unbiased estimator $\hat{v}_d(\mathbf{X})$ of parameter v^*_d, with variance-covariance matrix $Var(\hat{v}_d)$ satisfies $Var(\hat{v}_i) \succeq \mathcal{I}^{-1}(v^*_d)$, *i.e.*, the difference $Var(\hat{v}_d) - \mathcal{I}^{-1}(v^*_d)$ is positive semi-definite. As ML estimators are *asymptotically efficient*, this inequality becomes equality as the sample size grows to infinity (Everitt 2002). Therefore for a relatively large sample size, we can assume that the inverse of the Fisher information matrix is a good representative of variance of ML estimates, and by substituting the inverse of the variance matrix into Eq 6, we get an approximation to the combination using the inverse of variance matrix.

3. Analysis of NMR Spectra

Each pure chemical compound has a unique NMR "signature", which is a 1-dimensional signal composed of a set of clusters, where each cluster has a center and involves one or more peaks, each of which is characterized by 3 parameters, defining the peak's height a, width w and position z relative to the cluster center, within a Lorentzian function (Freeman 1987; Wishart et al. 2007). As these clusters do not move much, biochemists have long used NMR to determine the identity of a pure compound, based on the observed peak locations. Moreover, as the height of the peaks in a compound are essentially proportional to the concentration of that compound, they can also quantify that concentration. Hence, the NMR spectrum of a mixture of chemicals $\{c_m\}$ (appearing in, say, human blood or urine) is essentially just the linear combination of those signatures — $\sum_m \beta_m$ *signature*(c_m) where each β_m coefficient depends on the concentration of the associated compound c_ms — which potentially allows us to recover those concentrations from a mixture (Weljie et al. 2006). In fact, once we determine the centers $\alpha = (\alpha_p)$ for the clusters, we can then find the concentrations $\beta = (\beta_m)$ by the non-negative linear least square methods (Lawson and Hanson 1974). The challenge, however, is finding the cluster centers.

To be more precise, the following equation shows the 1-dimensional spectrum (over the discrete set of points on which spectrum is defined, $\mathcal{Y}$) produced by M metabolites, where the m^{th} metabolite has concentration $\beta_m \geq 0$ and involves clusters in the set $\Gamma(m)$, where the p^{th} cluster is at position α_p and involves peaks in $\Upsilon(p)$, each with its Lorentzian parameters $\langle\, w_r, a_r, d_r \,\rangle$: $\forall y \in \mathcal{Y} \quad S_{\alpha,\beta}(y) =$

$$\sum_{m=1}^{M} \beta_m \underbrace{\underbrace{\sum_{p\in\Gamma(m)} \underbrace{\sum_{r\in\Upsilon(j)} \underbrace{\frac{a_r w_r}{w_r + 4(\alpha_p + z_r - y)^2}}_{\text{peak, centered at } \alpha_p + z_r}}_{\text{cluster, centered at } \alpha_p}}_{\text{metabolite, with concentration } \beta_m}}} \qquad (7)$$

These parameter values and peak and cluster information appear in a predefined library that specify the NMR signatures of a set of compounds (Wishart et al. 2007).

Given an observed spectrum $\tilde{S}$, we want to find the

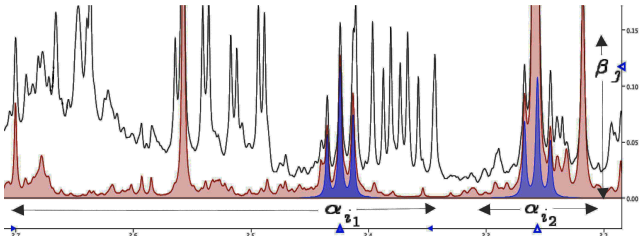

Figure 2: NMR spectrum, overlayed with two clusters α_{i_1} and α_{i_2}, associated with metabolite j. (Here $i_1, i_2 \in \Gamma(j)$.)

metabolite concentrations, $\beta = (\beta_m)$ over the set of possible compounds — which corresponds to $\{\hat{\alpha}, \hat{\beta}\} = \arg\min_{\alpha,\beta} L(\{\alpha, \beta\})$ where

$$L(\{\alpha, \beta\}) = \sum_{y \in \mathcal{Y}} [S_{\alpha,\beta}(y) - \tilde{S}(y)]^2 \quad (8)$$

is the L_2 error ; see Figure 2. (While we only care about the concentrations β, we must also determine the cluster centers α to find them.) If we knew the true concentrations β^*, we could measure the error using the average *controlled relative error*:

$$\kappa_1 \doteq \frac{1}{M} \sum_{m=1}^{M} \min\left(1, \frac{|\hat{\beta}_m - \beta^*_m|}{\beta^*_m}\right).$$

We also consider *relative absolute error* defined as:

$$\kappa_2 \doteq \frac{\sum_{m=1}^{M} |\hat{\beta}_m - \beta^*_m|}{\sum_{m=1}^{M} \beta^*_m}.$$

As we see in Eq 7, the value of the Lorentzian function drops quadratically with the distance $(\alpha_p + z_r - y)$ from the center of peak. We can therefore assume that each peak has a compact support, which consequently implies that each cluster (small set of close-by peaks) will potentially affect only a small region of the whole spectra. This region includes the bounded amount of shift in the center of the cluster α_p, and the effective width of cluster; call this region $\mathbb{Y}_p \subset \mathcal{Y}$. We can rewrite the optimization (Eq 8) as a weighted sum of squared error over all regions.

$$\{\hat{\alpha}, \hat{\beta}\} = \arg\min_{\alpha,\beta} \sum_{\substack{p \in \Gamma(m) \\ 1 \le m \le M}} L_p(\{\alpha, \beta\}_{\mathcal{M}(p)}) \quad (9)$$

where $\{\alpha, \beta\}_{\mathcal{M}(p)}$ is the set of variables appearing in sub-problem j. For convenience we define the function $\Lambda(p)$ that identifies the metabolite of cluster p: $\Lambda(p) = m$ iff $p \in \Gamma(m)$. Using this notation $\{\alpha, \beta\}_{\mathcal{M}(j)} = \{\{\alpha_{p'}, \beta_{\Lambda(p')}\}\}_{p' \in \mathcal{M}(p)}$ and $\mathcal{M}(p) = \{p' \ s.t. \ \mathbb{Y}_j \cap \mathbb{Y}_{p'} \neq \emptyset\}$. Now define each sub-problem as:

$$L_p(\{\alpha, \beta\}_{\mathcal{M}(p)}) = \sum_{y_k \in \mathbb{Y}_p} \eta_k \left(\tilde{S}(y_k) - \sum_{p' \in \mathcal{M}(j)} S_{\alpha_{p'}, \beta_{\Lambda(p')}}(y_k)\right)^2 \quad (10)$$

where

$$S_{\alpha_{p'}, \beta_{\Lambda(p')}}(y) = \beta_{\Lambda(p')} \sum_{r \in \Upsilon(p')} \frac{a_r w_r}{w_r + 4(z_r + \alpha_{p'} - y)^2}$$

is a single cluster that appears in sub-problem p.[3]

Figure 1(d) shows the coupling matrix for this optimization problem. Our CEED implementation uses Gaussian distributions, which means $v_d = [\mu_d, \sigma_d] \in \Re^2$. Therefore $\mathcal{N}_{v_{\mathcal{M}(p)}} = \mathcal{N}_{[\mu_{\mathcal{M}(p)}, \sigma_{\mathcal{M}(p)}]}$ is a distribution over the variables of p^{th} sub-problem, including both αs and βs.

Input: NMR spectrum $\tilde{S}(y)$; Library of NMR chemical signatures; Number of instances for sub-problem p at iteration t, $N(p,t)$; Elite fraction $\rho \in (0,1)$; Learning rate $\zeta \in (0,1)$; Stopping crition $t_{max} \in \mathcal{N}$; Max Standard Deviation σ_{max}.

Output: A vector $\mu^* = \{\hat{\alpha}^*, \hat{\beta}^*\}$ of concentration for metabolites and shift values for clusters.

1. For each cluster p in the library find $\mathbb{Y}_p$ — the area affected by each cluster — and using this derive sub-problems $L_p(.)$, $\mathcal{A}(d)$ and $\mathcal{M}(p)$. Set $t = 1$ and define an initial distribution f_{v^0} using $\mathbb{Y}_p$s and some upper-bound for concentrations.
2. For each sub-problem p, generate $N(p,t)$ instances $\{X^1, ..., X^{N(p,t)}\}$ from $\mathcal{N}_{v^{t-1}_{\mathcal{M}(p)}}$, where $N(p,t) \propto \sum_{p' \in \mathcal{M}(p)} \sigma^{t-1}_{p'}$. That is, the sample size for each sub-problem is proportional to the problem size/difficulty.
3. Evaluate the samples over each domain using Eq 10. Since our variables are bounded but the support of the associated Gaussian distributions is not, we may find some cluster center outside of its legal range; here we simply ignore any such instance. (Kroese, Porotsky, and Rubinstein (2006) use a similar approach in OCE.)
4. For each sub-problem p: find the elite instances (here using $\rho = 0.05$), and use them to estimate the maximum likelihood parameter $\hat{v}^t_{\mathcal{M}(p),p}$.
5. For each variable d, linearly combine its estimates $\{\hat{v}^t_{d,p}\}_{p \in \mathcal{A}(d)}$ using their Fisher information. For Gaussian distributions, the Fisher information is $\mathcal{I}([\mu, \sigma]) = \text{Diagonal}(\frac{1}{\sigma^2}, \frac{1}{2\sigma^4})$, from which we get the following linear combination:

$$\hat{\mu}^t_d = \frac{\sum_{p \in \mathcal{A}(d)} \frac{\hat{\mu}^t_{d,p}}{(\hat{\sigma}^t_{d,p})^2}}{\sum_{p \in \mathcal{A}(d)} \frac{1}{(\hat{\sigma}^t_{d,p})^2}} \quad \hat{\sigma}^t_d = \frac{\sum_{p \in \mathcal{A}(d)} \frac{1}{(\hat{\sigma}^t_{d,p})^2}}{\sum_{p \in \mathcal{A}(d)} \frac{1}{(\hat{\sigma}^t_{d,p})^4}} \quad (11)$$

[3] As a small technical issue: For $\hat{\alpha}, \hat{\beta}$ in Eq 9 to be the minimizers of Eq 8, we should avoid double-counting, perhaps by setting $\eta_k = |\{i : y_k \in \mathbb{Y}_i\}|^{-1}$. For example, for each $y_k \in \mathbb{Y}_{p_1} \cap \mathbb{Y}_{p_2} \cap \mathbb{Y}_{p_3}$, we could set $\eta_k = \frac{1}{3}$. However, for practical reasons we set all the weights η equal to 1 in our experiments. In the ideal case, as all loss functions L_p are equal to zero in global minima, any choice of weights gives the same zero error, and therefore this choice of $\eta \equiv 1$ will not matter. For real spectra, however, by removing the weights we are giving more weight to more critical regions, which produces better results in practice.

After this, we then update the parameter vector by setting $v^{t+1} = \zeta[\hat{\mu}^t, \hat{\sigma}^t] + (1-\zeta)v^t$, where $\zeta \in (0,1)$ is a learning rate.

6. If $\max(\sigma^{t+1}) < \sigma_{max}$ or $t \geq t_{max}$, return μ^{t+1} as the approximation to $\{\alpha^*, \beta^*\}$. Otherwise increase t, and return to Step 2.

In practice, the available metabolite library will not include all-and-only the metabolites in the chemical mixture — *i.e.*, there are some metabolites in the mixture that are not present in the library and vice versa. Here, CEED may produce an inferior fit by trying too hard to match the metabolites in its library. To reduce this problem we changed Step 5 to use the *minimum* of the different estimates of each concentration value (β_m), rather than their linear combination (Eq 11). We still use Eq 11 to combine α's. We also found that adding a term to the loss function of Eq 8 that is proportional to the *total variation* of the absolute error produces better fits for real spectra: $TV(\{\alpha, \beta\}) = \sum_{y \in \mathcal{Y} - \{0\}} |[S_{\alpha,\beta}(y) - \tilde{S}(y)] - [S_{\alpha,\beta}(y-1) - \tilde{S}(y-1)]|$. (*I.e.*, the loss function is a linear combination of Eq 8 and this TV term.) This term encourages CEED to produce a smooth difference (between the fit and the spectra), which helps CEED to avoid trying too hard to fit an region whose peaks correspond to compounds that are not present. This means our system will proprose a mixture over relatively few chemicals, which is appropriate, given the inherent incompleteness of our library.

Since CEED works with a decomposed loss function, we count the number of total evaluations of the Lorentzian function (peaks in Eq 7) as the measure of computation resource used by algorithms. We compared our algorithm against Gradient Descent (GD), Simulated Annealing (SA) and Genetic Algorithm (GA) methods on a typical simulated spectrum, over 90 metabolites involving 505 cluster centers (hence a total of 595 variables). We used the implementation of these methods provided by the standard Matlab™ toolbox.[4] Figure 3 compares the convergence rate of different algorithms for the typical spectrum, using the L_2 error (Eq 8).

We also applied our CEED system to a set of 39 manually-fitted 600 MHz real urine spectra, and compared our results to results obtained by SAGD (a hybrid of Simulated Annealing and Gradient Descent), a state-of-the-art tool provided by a company that is active in the analysis of NMR Spectra. Overall, our CEED achieved significantly better fits. Figure 4 shows CEED's fits. We see that CEED avoids excessively fitting available areas — that are results of baseline error and missing compounds — and also tends to produce smooth difference lines, which leads to better fits in general. We can also confirm the high quality of the fit by comparing the metabolite concentrations reported by CEED, to expert's estimates; see Table 1(left) which uses the κ_1 and κ_2 error measures defined above. We also considered the task of simply detecting the presence of a compound, by thresholding its reported concentration with the threshold of 0.02 mMol; see Table 1(right). Overall CEED performed better in both tasks. We use **bold-face** when CEED statistically significantly outperforms SAGD (paired t-test, $p <$5E-13).

[4] GD is implemented by constrained nonlinear optimization, which uses active-set and line-search. SA is using fast annealing with exponential temperature update; we report results based on the best reannealing interval. GA used ranking for fitness scaling, stochastic uniform method for parent selection, cross-over fraction of 0.8; we report the result for the best combination of population size and generations. All major choices are made by a reasonable effort of trial and error.

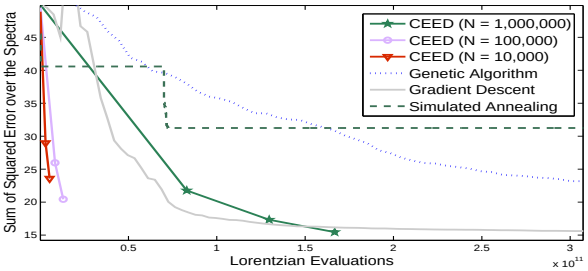

Figure 3: Comparing convergence rates on a simulated spectra. Horizontal axis: number of Lorentzian function evaluations. Vertical axis: L_2 norm of reconstruction error, Eq 8.

Table 1: Comparison of CEED and SAGD: (left) concentration errors and (right) presense detection.

Alg	κ_1	κ_2	Precision	Recall	F-measure
CEED	**.39** $\pm$ .05	**.43** $\pm$ .11	**.83** $\pm$ .08	.93 $\pm$ .06	**.87** $\pm$ .06
SAGD	.76 $\pm$ .05	.69 $\pm$ .17	.68 $\pm$ .13	.97 $\pm$ .03	.79 $\pm$ .10

4. Conclusion

This paper introduces a new stochastic optimization method CEED that attempts to find good solutions to partially decomposable problems, and demonstrates that it works effectively in the important real-world context of interpreting ^{1}H NMR spectra. Here, we found that CEED was considerable better than many existing optimization methods. (To demonstrate the generality of the method, we also applied CEED to several other decomposable problems, including SAT and Sudoku; see (RPG 2010).) This improvement is largely due to CEED's ability to use the sub-problem structure, which is clearly very important information. In addition to our theoretical claims, we have also produced a very practical system, one that can effectively analyse complex ^{1}H NMR spectra; this will prove extremely valuable in the study of metabolic bio-markers and helpful in diagnosing and treating diseases (Wishart et al. 2007).

Acknowledgement

We thank Drs. Vickie Baracos and David Wishart and their labs, and the researchers at Chenomx Inc. (especially Pascal Mercier and Jack Newton) for providing us with both data and useful insights.

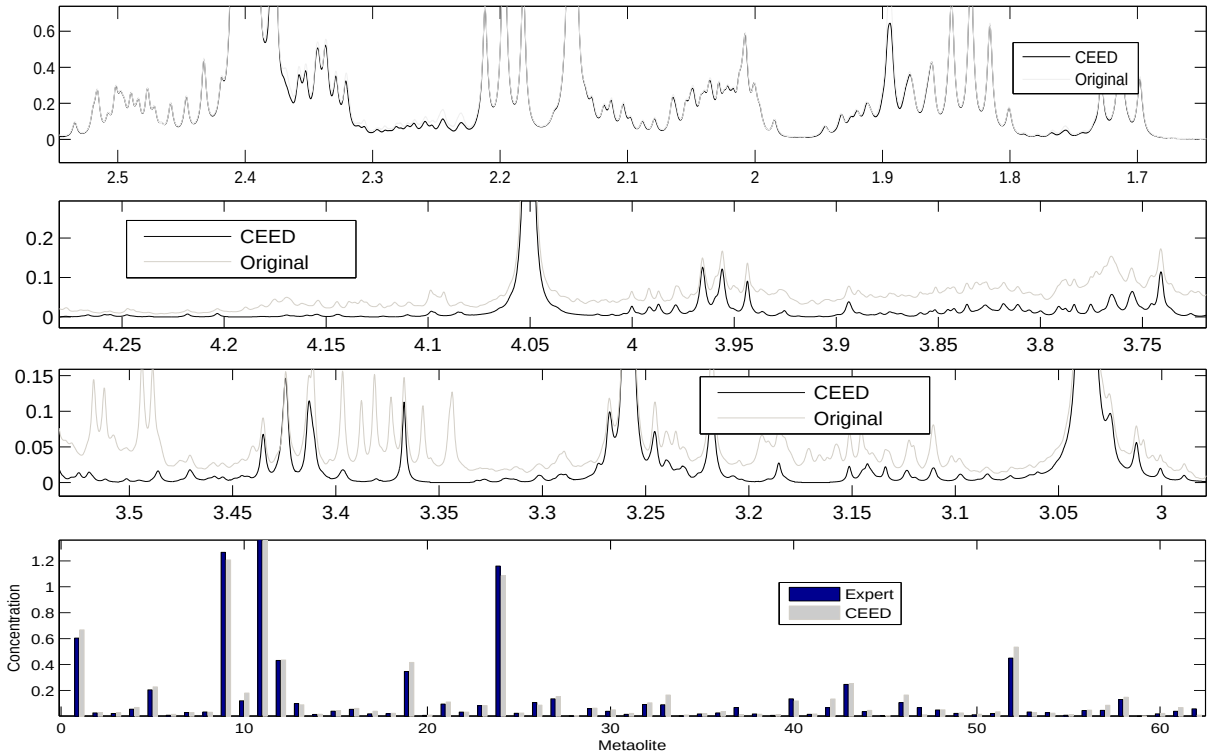

Figure 4: (top) CEED fit for a simulated spectrum; (middle) Two typical fits produced by CEED on real urine spectra; (bottom) Comparison of concentrations found by CEED with expert's estimate for a typical urine spectrum

References

Baluja, S. 2002. Using a priori knowledge to create probabilistic models for optimization. *J. Approximate Reasoning* 31:193–220.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Cramer, H. 1946. *Mathematical methods of statistics*. Princeton University Press, Princeton.

Everitt, B. 2002. *The Cambridge dictionary of statistics*. Cambridge University Press, 2nd edition.

Fisher, R. 1925. Theory of statistical estimation. *Cambridge Philosophical Society* 700–725.

Freeman, R. 1987. *A handbook of nuclear magnetic resonance*. Harlow Longman.

Garey, M., and Johnson, D. 1990. *Computers and Intractability*. W. H. Freeman and Co.

Kroese, D.; Porotsky, S.; and Rubinstein, R. 2006. The cross-entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability* 8:383–407.

Kullback, S. 1959. *Information theory and statistics*. John Wiley and Sons, NY.

Lawson, C., and Hanson, R. 1974. *Solving Least Squares Problems*. Prentice-Hall.

Lehmann, E., and Casella, G. 1998. *Theory of Point Estimation*. Springer-Verlag.

Rao, C. 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37:81–91.

RPG. 2010. `http://sites.google.com/site/cefornmr/`.

Rubinstein, R., and Kroese, D. 2004. *The Cross-Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer.

Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; and Slupsky, C. M. 2006. Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data. *Anal Chem* 78(13):4430–42.

Wishart, D.; Tzur, D.; Knox, C.; Eisner, R.; and Guo, A. 2007. HMDB: the human metabolome database. *Nucleic acids research*.