

Minimum Error Rate Training and the Convex Hull Semiring*

Chris Dyer

School of Computer Science
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA
cdyer@cs.cmu.edu

Abstract

We describe the line search used in the minimum error rate training algorithm (Och, 2003) as the “inside score” of a weighted proof forest under a semiring defined in terms of well-understood operations from computational geometry. This conception leads to a straightforward complexity analysis of the dynamic programming MERT algorithms of Macherey et al. (2008) and Kumar et al. (2009) and practical approaches to implementation.

1 Introduction

Och’s (2003) algorithm for minimum error rate training (MERT) is widely used in the direct loss minimization of linear translation models. It is based on an efficient and optimal line search and can optimize non-differentiable, corpus-level loss functions. While the original algorithm used n -best hypothesis lists to learn from, more recent work has developed dynamic programming variants that leverage much larger sets of hypotheses encoded in finite-state lattices and context-free hypergraphs (Macherey et al., 2008; Kumar et al., 2009; Sokolov and Yvon, 2011). Although MERT has several attractive properties (§2) and is widely used in MT, previous work has failed to explicate its close relationship to more familiar inference algorithms, and, as a result, it is less well understood than many other optimization algorithms.

*While preparing these notes, I discovered the work by Sokolov and Yvon (2011), who elucidated the semiring properties of the MERT line search computation. Because they did not discuss the polynomial bounds on the growth of the values while running the inside algorithm, I have posted this as an unpublished manuscript.

In this paper, we show that the both the original (Och, 2003) and newer dynamic programming algorithms given by Macherey et al. (2008) and Kumar et al. (2009) can be understood as weighted logical deductions (Goodman, 1999; Lopez, 2009; Eisner and Filardo, 2011) using weights from a previously undescribed semiring, which we call the **convex hull semiring** (§3). Our description of the algorithm in terms of semiring computations has both theoretical and practical benefits: we are able to provide a straightforward complexity analysis and an improved DP algorithm with better asymptotic and observed run-time (§4). More practically still, since many tools for structured prediction over discrete sequences support generic semiring-weighted inference (Allauzen et al., 2007; Li et al., 2009; Dyer et al., 2010; Eisner and Filardo, 2011), our analysis makes it possible to add dynamic programming MERT to them with little effort.

2 Minimum error rate training

The goal of MERT is to find a weight vector $\mathbf{w}^* \in \mathbb{R}^d$ that minimizes a corpus-level loss \mathcal{L} (with respect to a development set \mathcal{D}) incurred by a decoder that selects the most highly-weighted output of a linear structured prediction model parameterized by feature vector function \mathbf{H} :

$$\begin{aligned} \mathbf{w}^* &= \arg \min_{\mathbf{w}} \mathcal{L}(\{\hat{\mathbf{y}}_i^{\mathbf{w}}\}, \mathcal{D}) \\ \{\hat{\mathbf{y}}_i^{\mathbf{w}}\} &= \arg \max_{\mathbf{y} \in \mathcal{Y}(x_i)} \mathbf{w}^\top \mathbf{H}(x_i, \mathbf{y}) \quad \forall (x_i, \mathbf{y}_i^{\text{gold}}) \in \mathcal{D} \end{aligned}$$

We assume that the loss \mathcal{L} is computed using a vector error count function $\delta(\hat{y}, y) \rightarrow \mathbb{R}^m$ and a loss

scalarizer $L : \mathbb{R}^m \rightarrow \mathbb{R}$, and that the error count decomposes linearly across examples:¹

$$\mathcal{L}(\{\hat{\mathbf{y}}_i^{\mathbf{w}}\}, \mathcal{D}) = L \left(\sum_{i=1}^{|\mathcal{D}|} \delta(\hat{\mathbf{y}}_i, \mathbf{y}_i^{\text{gold}}) \right)$$

At each iteration of the optimization algorithm, MERT choses a starting weight vector \mathbf{w}_0 and a search direction vector \mathbf{v} (both $\in \mathbb{R}^d$) and determines which candidate in a set has the highest model score for *all* weight vectors $\mathbf{w}' = \eta\mathbf{v} + \mathbf{w}_0$, as η sweeps from $-\infty$ to $+\infty$.²

To understand why this is potentially tractable, consider any (finite) set of outputs $\{\mathbf{y}_j\} \subseteq \mathcal{Y}(x)$ for an input x (e.g., an n -best list, a list of n random samples, or the complete proof forest of a weighted deduction). Each output \mathbf{y}_j has a corresponding feature vector $\mathbf{H}(x, \mathbf{y}_j)$, which means that the *model score* for each hypothesis, together with η , form a line in \mathbb{R}^2 :

$$\begin{aligned} s(\eta) &= (\eta\mathbf{v} + \mathbf{w}_0)^\top \mathbf{H}(x, \mathbf{y}_j) \\ &= \underbrace{\eta \mathbf{v}^\top \mathbf{H}(x, \mathbf{y}_j)}_{\text{slope}} + \underbrace{\mathbf{w}_0^\top \mathbf{H}(x, \mathbf{y}_j)}_{y\text{-intercept}}. \end{aligned}$$

The upper part of Figure 1 illustrates how the model scores (y -axis) of each output in an example hypothesis set vary with η (x -axis). The lower part shows how this induces a piecewise constant error surface (i.e., $\delta(\hat{\mathbf{y}}^{\eta\mathbf{v}+\mathbf{w}_0}, \mathbf{y}^{\text{gold}})$). Note that \mathbf{y}_3 has a model score that is always strictly less than the score of some other output at all values of η . Detecting such “obscured” lines is useful because it is unnecessary to compute their error counts. There is simply no setting of η that will yield weights for which \mathbf{y}_3 will be ranked highest by the decoder.³

¹Nearly every evaluation metric used in NLP and MT fulfills these criteria, including F-measure, BLEU, METEOR, TER, AER, and WER. Unlike many dynamic programming optimization algorithms, the error count function δ is not required to decompose with the structure of the model.

²Several strategies have been proposed for selecting \mathbf{v} and \mathbf{w}_0 . For an overview, refer to Galley and Quirk (2011) and references therein.

³Since δ need only be evaluated for the (often small) subset of candidates that can obtain the highest model score at some η , it is possible to use relatively computationally expensive loss

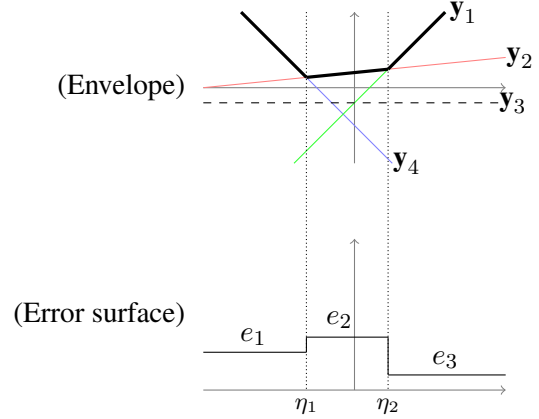


Figure 1: The model scores of a set of four output hypotheses $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}$ under a linear model with parameters $\mathbf{w} = \eta\mathbf{v} + \mathbf{w}_0$, inducing segments $(-\infty, \eta_1], [\eta_1, \eta_2], [\eta_2, \infty)$, which correspond (below) to error counts e_1, e_2, e_3 .

By summing the error surfaces for each sentence in the development set, a *corpus-level* error surface is created. Then, by traversing this from left to right and selecting best scoring segment (transforming each segment’s corpus level error count to a loss with L), the optimal η for updating \mathbf{w}_0 can be determined.⁴

2.1 Point-line duality

The set of line segments corresponding to the maximum model score at every η form an *upper envelope*. To determine which lines (and corresponding hypotheses) these are, we turn to standard algorithms from computational geometry. While algorithms for directly computing the upper envelope of a set of lines do exist, we proceed by noting that computing the upper envelope has as a dual problem that can be solved instead: finding the lower *convex hull* of a set of points (de Berg et al., 2010). The dual representation of a line of the form $y = mx + b$ is the point $(m, -b)$. This, for a given output, \mathbf{w}_0 , \mathbf{v} , and feature vector \mathbf{H} , the line showing how the model score of the output hypothesis varies with η can simply be represented by the point $(\mathbf{v}^\top \mathbf{H}, -\mathbf{w}_0^\top \mathbf{H})$.

functions. Zaidan and Callison-Burch (2009) exploit this and find that it is even feasible to solicit human judgments while evaluating δ !

⁴Macherey et al. (2008) recommend selecting the midpoint of the segment with the best loss, but Cer et al. (2008) suggest other strategies.

Figure 2 illustrates the line-point duality and the relationship between the primal upper envelope and dual lower convex hull. Usefully, the η coordinates (along the x -axis in the primal form) where upper-envelope lines intersect and the error count changes are simply the *slopes* of the lines connecting the corresponding points in the dual.

3 The Convex Hull Semiring

Definition 1. A *semiring* K is a quintuple $\langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$ consisting of a set \mathbb{K} , an addition operator \oplus that is associative and commutative, a multiplication operator \otimes that is associative, and the values $\bar{0}$ and $\bar{1}$ in \mathbb{K} , which are the additive and multiplicative identities, respectively. \otimes must distribute over \oplus from the left or right (or both), i.e., $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ or $(b \oplus c) \otimes a = (b \otimes a) \oplus (c \otimes a)$. Additionally, $\bar{0} \otimes u = \bar{0}$ must hold for any $u \in \mathbb{K}$. If a semiring K has a commutative \otimes operator, the semiring is said to be commutative. If K has an idempotent \oplus operator (i.e., $a \oplus a = a$ for all $a \in \mathbb{K}$), then K is said to be idempotent.

Definition 2. The Convex Hull Semiring. Let $(\mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1})$ be defined as follows:

| | |
|---------------|--|
| \mathbb{K} | A set of points in the plane that are the extreme points of a convex hull. |
| $A \oplus B$ | $\text{conv}[A \cup B]$ |
| $A \otimes B$ | convex hull of the Minkowski sum, i.e., $\text{conv}\{(a_1 + b_1, a_2 + b_2) \mid (a_1, a_2) \in A \wedge (b_1, b_2) \in B\}$ |
| $\bar{0}$ | \emptyset |
| $\bar{1}$ | $\{(0, 0)\}$ |

Theorem 1. The Convex Hull Semiring fulfills the semiring axioms and is commutative and idempotent.

Proof. To show that this is a semiring, we need only to demonstrate that commutativity and associativity hold for both addition and multiplication, from which distributivity follows. Commutativity ($A \cdot B = B \cdot A$) follows straightforwardly from the definitions of addition and multiplication, as do the identities. Proving associativity is a bit more subtle on account of the conv operator. For multiplication, we rely on results of Krein and Šmulian (1940), who show that

$$\text{conv}[A +_{\text{Mink.}} B] = \text{conv}[\text{conv } A +_{\text{Mink.}} \text{conv } B] .$$

For addition, we make an informal argument that a context hull circumscribes a set of points, and convexification removes the interior ones. Thus, addition continually expands the circumscribed sets, regardless of what their interiors were, so order does not matter. Finally, addition is idempotent since $\text{conv}[A \cup A] = A$. \square

4 Complexity

Shared structures such as finite-state automata and context-free grammars encode an exponential number of different derivations in polynomial space. Since the values of the convex hull semiring are themselves sets, it is important to understand how their sizes grow. Fortunately, we can state the following tight bounds, which guarantee that growth will be worst case linear in the size of the input grammar:

Theorem 2. $|A \oplus B| \leq |A| + |B|$.

Theorem 3. $|A \otimes B| \leq |A| + |B|$.

The latter fact is particularly surprising, since multiplication appears to have a bound of $|A| \times |B|$. The linear (rather than multiplicative) complexity bound for Minkowski addition is the result of Theorem 13.5 in de Berg et al. (2010). From these inequalities, it follows straightforwardly that the number of points in a derivation forest's total convex hull is upper bounded by $|E|$.⁵

Acknowledgements

We thank David Mount for suggesting the point-line duality and pointing us to the relevant literature in computational geometry and Adam Lopez for the TikZ MERT figures.

References

- [Allauzen et al.2007] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Proc. of CIAA*, volume 4783 of *Lecture Notes in Computer Science*. Springer. <http://www.openfst.org>.
- [Cer et al.2008] D. Cer, D. Jurafsky, and C. D. Manning. 2008. Regularization and search for minimum error rate training. In *Proc. ACL*.

⁵This result is also proved for the lattice case by Macherey et al. (2008).

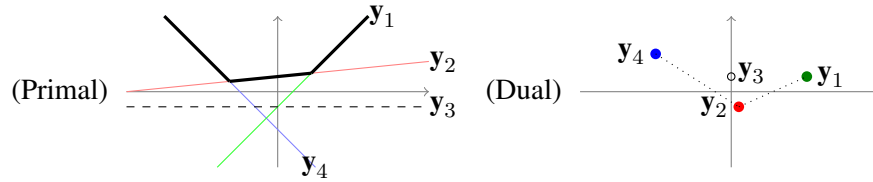


Figure 2: Primal and dual forms of a set of lines. The upper envelope is shown with heavy line segments in the primal form. In the dual, primal lines are represented as points, with upper envelope lines corresponding to points on the lower convex hull. The dashed line y_3 is obscured from above by the upper envelope in the primal and (equivalently) lies above the lower convex hull of the dual point set.

- [de Berg et al.2010] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. 2010. *Computational Geometry: Algorithms and Applications*. Springer, third edition.
- [Dyer et al.2010] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. of ACL*.
- [Eisner and Filardo2011] J. Eisner and N. W. Filardo, 2011. *Datalog 2.0*, chapter Dyna: Extending Datalog For Modern AI. Springer.
- [Galley and Quirk2011] M. Galley and C. Quirk. 2011. Optimal search for minimum error rate training. In *Proc. of EMNLP*.
- [Goodman1999] J. Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25(4):573–605.
- [Krein and Šmulian1940] M. Krein and W. Šmulian. 1940. On regularly convex sets in the space conjugate to a Banach space. *Annals of Mathematics (2)*, Second series, 41(3):556–583.
- [Kumar et al.2009] S. Kumar, W. Macherey, C. Dyer, and F. Och. 2009. Efficient minimum error rate training and minimum Bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of ACL-IJCNLP*.
- [Li et al.2009] Z. Li, C. Callison-Burch, C. Dyer, S. Khudanpur, L. Schwartz, W. Thornton, J. Weese, and O. Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proc. of the Fourth Workshop on Statistical Machine Translation*.
- [Lopez2009] A. Lopez. 2009. Translation as weighted deduction. In *Proc. of EACL*.
- [Macherey et al.2008] W. Macherey, F. J. Och, I. Thayer, and J. Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proc. of EMNLP*.
- [Och2003] F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*.
- [Sokolov and Yvon2011] A. Sokolov and F. Yvon. 2011. Minimum error rate training semiring. In *Proc. of AMTA*.
- [Zaidan and Callison-Burch2009] O. F. Zaidan and C. Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proc. of EMNLP*.