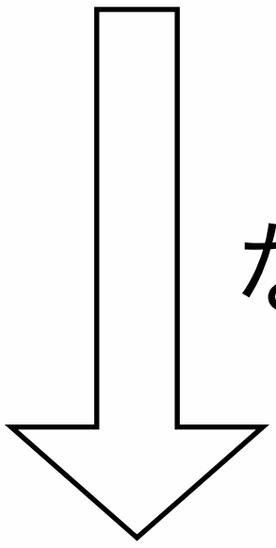
オンライン学習

渡辺太郎

taro.watanabe at nict.go.jp

https://sites.google.com/site/alaginmt2014/

機械翻訳について勉強したい。



なるべく良い翻訳...

I want to study about machine translation.
I need to master machine translation.
machine translation want to study.

infobox ）

infobox buddhist
道元 は 鎌倉 時代 初期 の で あ る 。

道元 （ どう げん ） は 、 鎌倉 時代 初期 の 禅僧 。

曹洞 禅 の 開祖

曹洞 宗 の 開祖 。

その 生涯 に は kigen 名 が あ る 。

晩年 に 希玄 と い う 異称 も 用い た 。

一般 に は 宗 と 呼 ば れ る こと に よ っ て 尊称 は 高僧 が あ る 。

同宗旨 で は 高祖 と 尊称 さ れ る 。

死後 に と い っ た 仏所 伝灯 国師 、 joyo-daishi で あ る 。

諡 は 、 仏性 伝東 国師 、 承陽 大師_ (僧) 。

一般 に は 道元 禅師 と 呼 ば れ る 。

一般 に は 道元 禅師 と 呼 ば れ る 。

また ～ 686 の 普及 に つ い て の 修行 を tooth brushing 、 大峰 、 食事 作法 cleaning と し て 、 日本 に お い て い る 。

日本 に 歯磨き 洗面 、 食事 の 際 の 作法 や 掃除 の 習慣 を 広め た と い わ れ る 。

エラー

- 探索エラー: スコアの高い翻訳を出すのに失敗
- モデルエラー: スコアの高い翻訳が誤っている
- 学習データの問題: 小さい、異なる
- 手軽な対処: 最適化(チューニング)

どうしましょう?

f = 機械翻訳について勉強したい

	$\log Pr(\phi\|e)$	$\log Pr(e)$	$\log Pr(f,\alpha\|\phi)$	
I want to study about machine translation	-2	-3	-4	-9
I need to master machine translation	-3	-4	-4	-11
machine translation want to study	-2	-5	-1	-8
I don't want to learn anything	-5	-2	-3	-10

0.5×-2	0.4×-3	0.2×-4	-3.0
0.5×-3	0.4×-4	0.2×-4	-3.9
0.5×-2	0.4×-5	0.2×-1	-3.2
0.5×-5	0.4×-2	0.2×-3	-3.9

kbestを正しく並び替えるように重みを学習

重み付け

$$\hat{e} = \arg\max_{e} Pr(f, \alpha|\phi, e)^{0.2}\ Pr(\phi|e)^{0.5}\ Pr(e)^{0.4}$$

- より一般化:

$$\begin{aligned}\hat{e} &= \arg\max_{e} \frac{\sum_{d} \exp\left(w^\top h(f, d, e)\right)}{\sum_{e', d'} \exp\left(w^\top h(f, d', e')\right)} \\ &\approx \arg\max_{\langle e, d\rangle} w^\top h(f, d, e)\end{aligned}$$

- 複数の素性h(e,d, f)をlog-linearに組み合わせ

最適化 = 最適なwを学習

MTパイプライン

kyoto-train.{ja,en}　対訳データ　大量(低品質?)

翻訳モデル　言語モデル

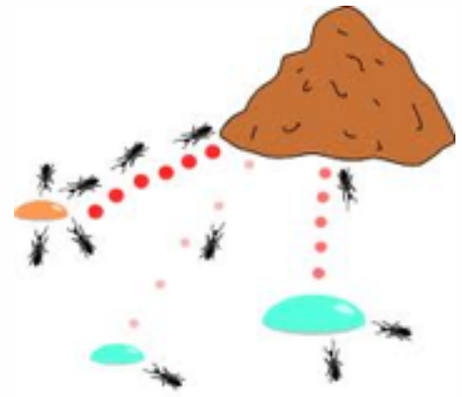
デコーダ

最適化の問題

重みパラメータ

(少量?)高品質　対訳データ　kyoto-tune.{ja,en}

最適化



$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}_{Pr(F,E)}[\ell(F, E; \boldsymbol{w})]$$

$$= \arg\min_{\boldsymbol{w}\in\mathcal{W}} \ell(F, E; \boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$$

- ある損失関数 ℓ を仮定し、対訳データ(F, E)に対するリスクを最小化□□
- 真の分布は未知なため、正則化 (Ω) された経験リスクを最小化

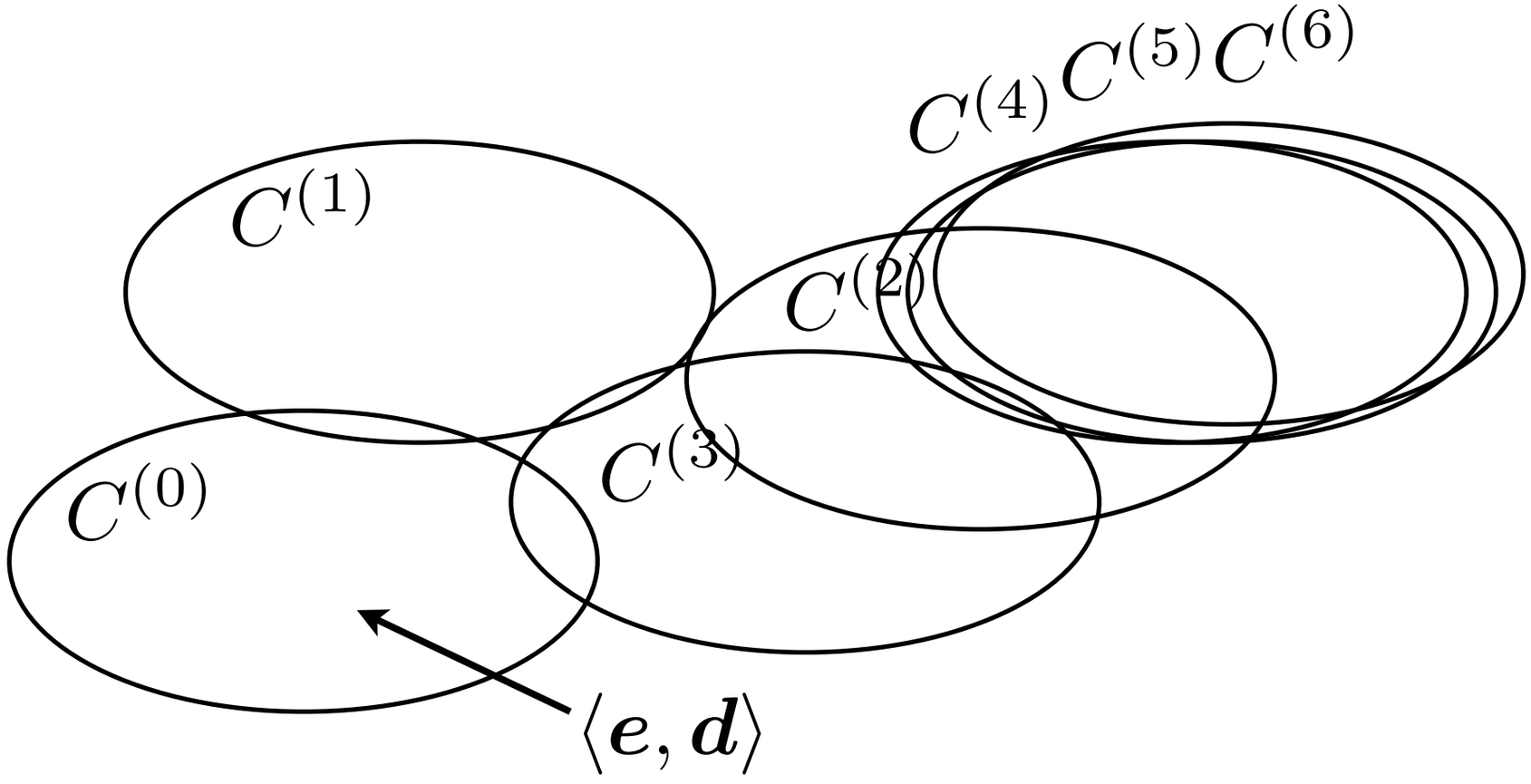
kbest近似最適化

1: **procedure** LEARN($\langle F, E \rangle$)
2: $C = \{\}$
3: **for** $t = 1...T$ **do**
4: $\boldsymbol{w}$でデコード、kbestリストを生成
5: kbestリストをCへ結合
6: $\boldsymbol{w}^{(t+1)} = \arg\min_{\boldsymbol{w}} \ell(F, E, C; \boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$
7: **end for**
8: **end procedure**

(Och and Ney, 2002)

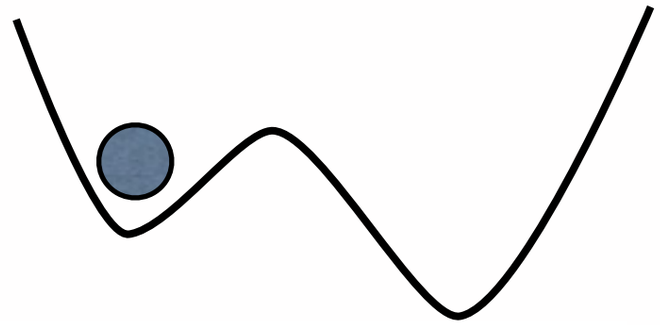
- “k-best結合”以外の近似?

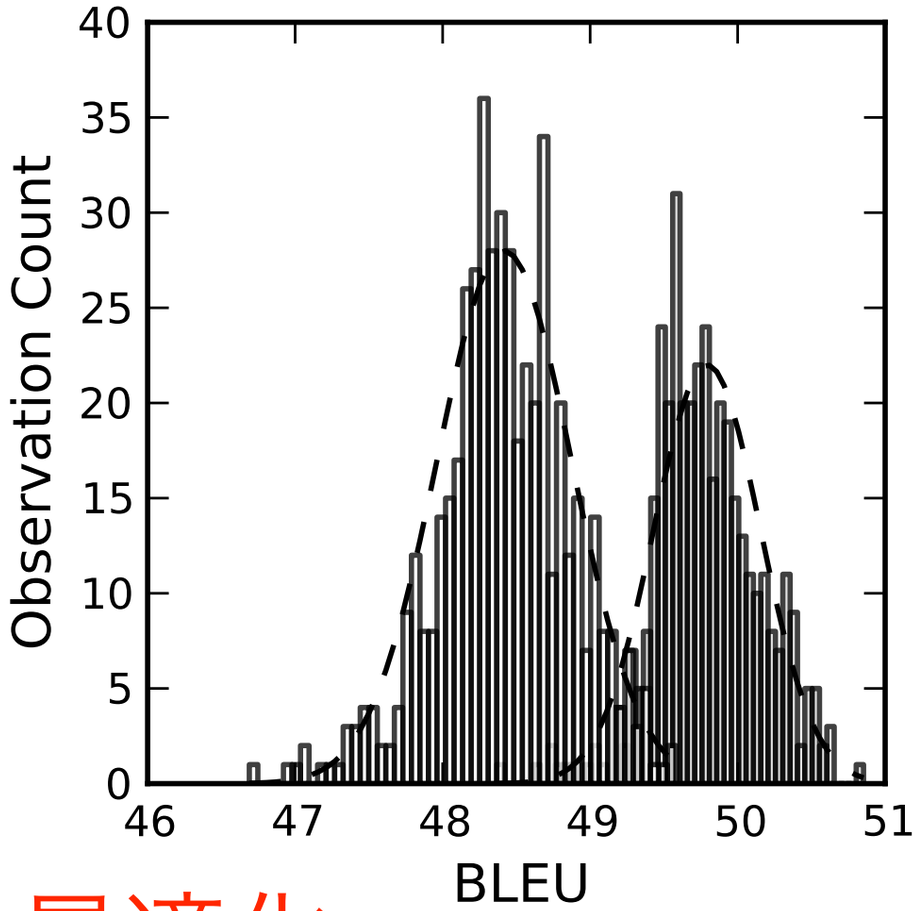
kbest近似最適化

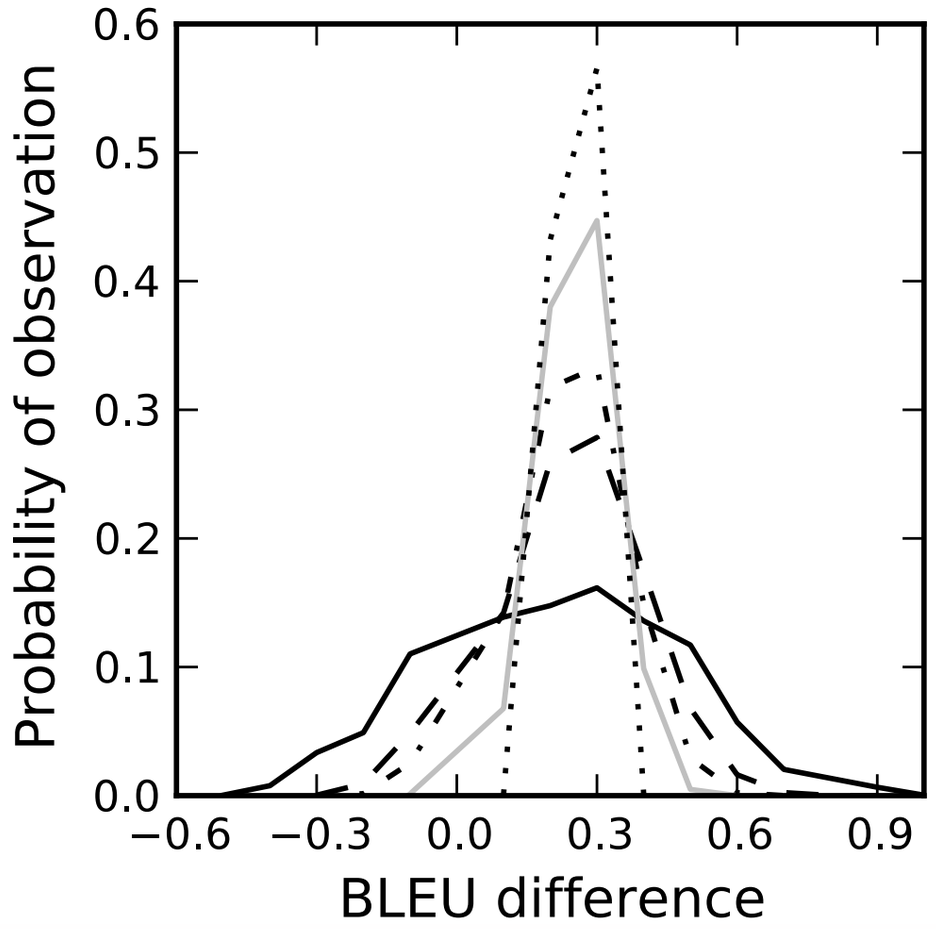


- 各繰り返しでk-bestを結合しつつ学習
- 収束はするけど.... 局所解

局所解







(Clark et al., 2011)

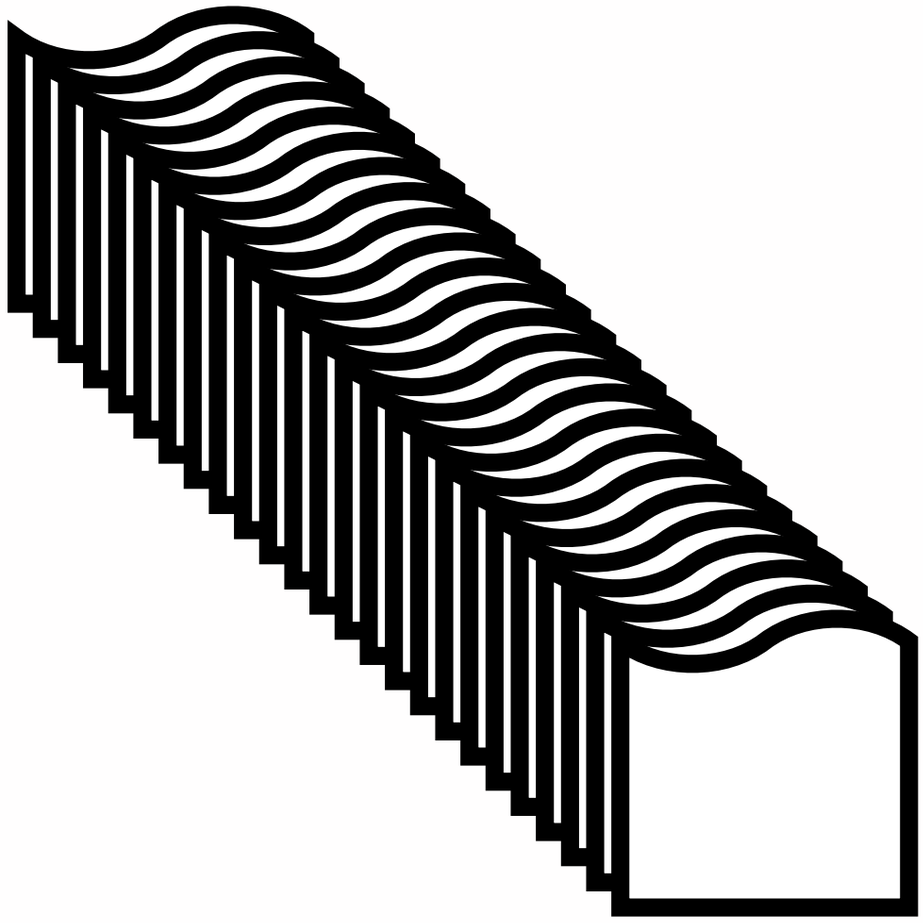
最適化

- ~~MERT~~は、パラメータの初期値に大きく依存

最適化

- ランダムな初期値から複数~~MERT~~

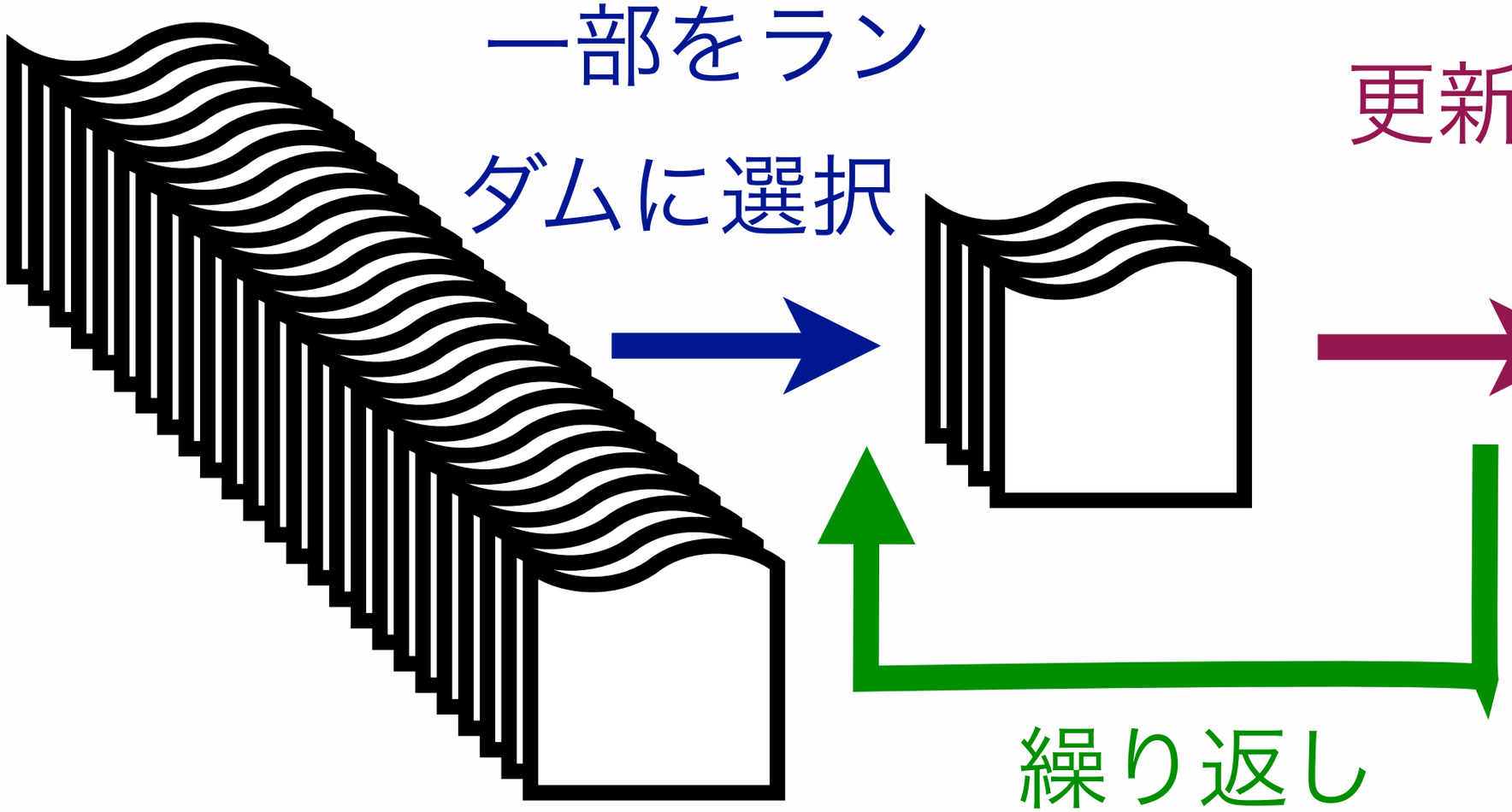
大規模化の問題



$$h(f, d, e) = \begin{pmatrix} \vdots \\ \cdots \\ \vdots \\ \cdots \\ \vdots \\ \cdots \\ \vdots \\ \cdots \\ \vdots \end{pmatrix}$$

- データが多い、高次元な素性ベクトル
- 効率のよい学習手法?

どうせ近似するなら…



- 学習データを近似
- 非常に簡単なアルゴリズムで実現

オンライン学習

1: **procedure** ONLINELEARN($\langle F, E \rangle = \left\{ \langle \boldsymbol{f}^{(i)}, \boldsymbol{e}^{(i)} \rangle \right\}_{i=1}^{N}$)
2: $\boldsymbol{w}^{(1)} \leftarrow \emptyset$
3: **for** $t \in \{1, \ldots, T\}$ **do**
4: $\langle \tilde{F}^{(t)}, \tilde{E}^{(t)} \rangle \subseteq \langle F, E \rangle$
5: $\tilde{C}^{(t)} \leftarrow \mathrm{GEN}(\tilde{F}^{(j)}, \boldsymbol{w}^{(t)})$
6: $\boldsymbol{w}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{w} \in \mathcal{W}} \ell(\tilde{F}^{(t)}, \tilde{E}^{(t)}, \tilde{C}^{(t)}; \boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$
7: **end for**
8: **return** $\boldsymbol{w}^{(T+1)}$
9: **end procedure**

ランダムにサンプリング: mini batch

デコード

更新

- 自然言語処理ではよく使われる: 形態素解析、構文解析etc.
- 当然、機械翻訳にも使えます(Watanabe et al., 2007)

問題: BLEU

e_1 I want to study about machine translation

e_2 I need to master machine translation

e_3 machine translation want to study

e_4 I don't want to learn anything

- k-bestのランキングに基づく最適化
- コーパス単位のBLEU ≠文単位のBLEUの平均
- 文単位では、正しく学習するのは困難

BLEUの近似

$$\mathsf{GEN}(\boldsymbol{f}^{(s)}, \boldsymbol{w})$$

$$\boldsymbol{e}^{(1)}, \cdots, \begin{pmatrix} \boldsymbol{c}_1^{(s)} \\ \vdots \\ \boldsymbol{c}_i^{(s)} \\ \vdots \\ \boldsymbol{c}_K^{(s)} \end{pmatrix}, \cdots, \boldsymbol{e}^{(S)}$$

- 今までの各文に対するBLEUの統計量を保存(1-bestあるいはoracle) (Watanabe et al., 2007)

減衰によるBLEUの近似

$$\begin{aligned} \boldsymbol{b} &\leftarrow 0.9 \times (\boldsymbol{b} + \boldsymbol{c}(\boldsymbol{e})) \\ l &\leftarrow 0.9 \times (l + |\boldsymbol{f}|) \end{aligned}$$

$$B(\boldsymbol{e}) = (l + |\boldsymbol{f}|) \times \mathsf{Bleu}(\boldsymbol{b} + \boldsymbol{c}(\boldsymbol{e}))$$

$$\begin{aligned} \hat{\boldsymbol{e}}^{(s)} &= \operatorname*{argmax}_{\boldsymbol{e}} -B(\boldsymbol{e}) + \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}^{(s)}, \boldsymbol{e}) \\ \dot{\boldsymbol{e}}^{(s)} &= \operatorname*{argmax}_{\boldsymbol{e}} +B(\boldsymbol{e}) + \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}^{(s)}, \boldsymbol{e}) \end{aligned}$$

- sentence-BLEUに対して、今までのBLEUの履歴(×0.9)を加える(Chiang et al., 2008)

パーセプトロン

procedure PERCEPTRON($\langle F, E \rangle = \left\{ \langle \boldsymbol{f}^{(i)}, \boldsymbol{e}^{(i)} \rangle \right\}_{i=1}^{N}$)

  $\boldsymbol{w}^{(1)} \leftarrow \mathbf{0}$

  for $t \in \{1 \ldots T\}$ **do**

    $\langle \boldsymbol{f}, \boldsymbol{e} \rangle \sim \langle F, E \rangle$

    $\tilde{\boldsymbol{c}} \leftarrow \mathrm{GEN}(\boldsymbol{f}, \boldsymbol{w}^{(t)})$

    $\langle \hat{\boldsymbol{e}}, \hat{\boldsymbol{d}} \rangle \in \tilde{\boldsymbol{c}}$

    $\langle \boldsymbol{e}^*, \boldsymbol{d}^* \rangle \in \tilde{\boldsymbol{o}} \subseteq \tilde{\boldsymbol{c}}$

    if $\langle \boldsymbol{e}^*, \boldsymbol{d}^* \rangle \neq \langle \hat{\boldsymbol{e}}, \hat{\boldsymbol{d}} \rangle$ **then**

      $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}^*, \boldsymbol{d}^*) - \boldsymbol{h}(\boldsymbol{f}, \hat{\boldsymbol{e}}, \hat{\boldsymbol{d}})$

    end if

  end for

  return $\boldsymbol{w}^{(T+1)}$あるいは$\frac{1}{T}\sum_{t=2}^{T+1} \boldsymbol{w}^{(t)}$

end procedure

一文だけサンプル

1-best

オラクル翻訳

間違ったらペナルティ

最後のパラメータor平均値

パーセプトロン

e_3 machine translation want to study

e_2 I need to master machine translation

e_1 I want to study about machine translation

e_4 I don't want to learn anything

$$\arg\min_{\boldsymbol{w}} \max\left\{0, 0 - \left(\boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, e_1) - \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, e_3)\right)\right\}$$

- e_1のスコアが、e_3のスコアよりも、大きくなるように、wを更新

hinge損失

e_3 machine translation want to study

e_2 I need to master machine translation

e_1 I want to study about machine translation

e_4 I don't want to learn anything

$$\arg\min_{\boldsymbol{w}} \max\left\{0, 1 - \left(\boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1) - \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_3)\right)\right\}$$

- e_1のスコアが、e_3のスコアよりも、1 以上大きくなるように、wを更新

MIRA

$$\arg\min_{\boldsymbol{w}} \lambda \frac{1}{2} \|\boldsymbol{w} - \boldsymbol{w}^{(t)}\|^2 + \Delta\,\mathrm{error}(\boldsymbol{e}_1, \boldsymbol{e}_3) - \left(\boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1) - \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_3)\right)$$

(Crammer et al., 2006)

- 以前のパラメータ$w^{(t)}$との差分を小さくしつつ、 e_1のスコアが、e_3のスコアよりも、翻訳のエラー以上大きくなるように、wを更新

MIRA

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + \alpha^{(t)} \Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_3)$$

$$\alpha^{(t)} = \min \left\{ \frac{1}{\lambda}, \frac{\Delta \,\mathrm{error}(\boldsymbol{e}_1, \boldsymbol{e}_3) - \boldsymbol{w}^{(t)\top} \Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_3)}{\| \Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_3) \|^2} \right\}$$

$$\Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_3) = \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1) - \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_3)$$

- 非常に簡単な更新: $\alpha^{(t)}$により更新の量を調整
- $\alpha^{(t)}=1$: パーセプトロン

  他にも、CWやAROWなど様々な更新式

大きいマージン

e_3 machine translation want to study

e_2 I need to master machine translation

e_1 I want to study about machine translation

e_4 I don't want to learn anything

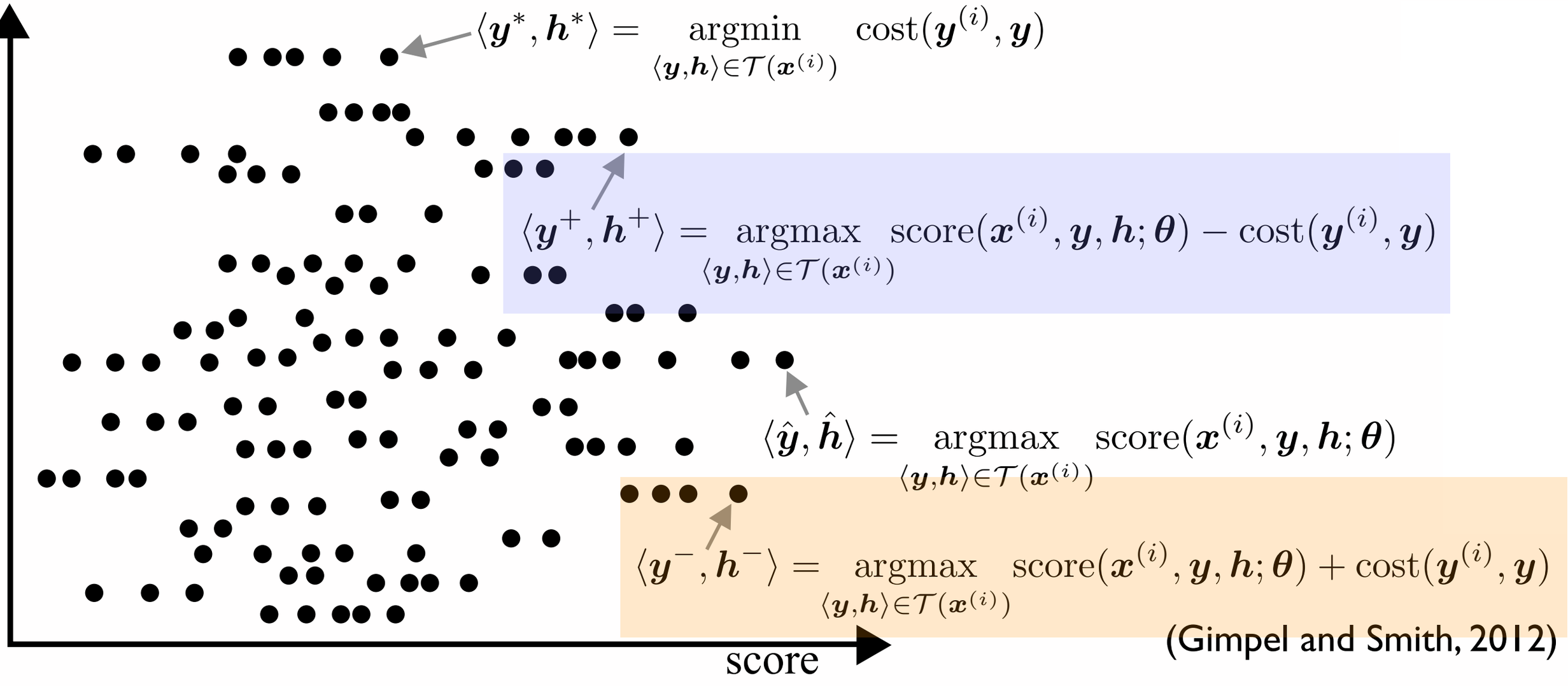
$$\Delta\,\mathrm{error}(\boldsymbol{e}_1, \boldsymbol{e}_3) - \boldsymbol{w}^{(t)\top} \Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_3)$$

あるいは

$$\Delta\,\mathrm{error}(\boldsymbol{e}_2, \boldsymbol{e}_3) - \boldsymbol{w}^{(t)\top} \Delta \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_2, \boldsymbol{e}_3)$$

- 正しい翻訳と誤った翻訳のうち、誤りの差が大きく、かつ、スコアの差が大きいペアを選択

(Chiang et al., 2008; Chiang et al., 2009)



(Gimpel and Smith, 2012)

SGD

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta^{(t+1)} \left(\Delta \ell(\tilde{F}^{(t)}, \tilde{E}^{(t)}, \tilde{C}^{(t)}; \boldsymbol{w}) + \lambda \Delta \Omega(\boldsymbol{w}) \right)$$

- 勾配が計算できるなら、hinge損失以外でも使える
- 学習率ηは、学習が進むにつれて、減衰

AdaGrad

$$\boldsymbol{g}^{(t+1)} \leftarrow \boldsymbol{g}^{(t)} + \Delta\ell(\tilde{F}^{(t)}, \tilde{E}^{(t)}, \tilde{C}^{(t)}; \boldsymbol{w})^2$$

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \frac{\eta_0}{\sqrt{\boldsymbol{g}^{(t+1)}}} \left(\Delta\ell(\tilde{F}^{(t)}, \tilde{E}^{(t)}, \tilde{C}^{(t)}; \boldsymbol{w}) + \lambda\Delta\Omega(\boldsymbol{w}) \right)$$

(Duchi et al., 2011)

- 各素性毎に、学習率を自動的に調整
- 他にも、RDA、AdaDec、AdaDeltaなど

並列化

1: **procedure** ParallelLearn($\langle F, E \rangle$)
2: $\{\langle F_1, E_1 \rangle, \ldots, \langle F_S, E_S \rangle\} \leftarrow$ Split($\langle F, E \rangle$)
3: **for** $s \in \{1, \ldots, S\}$ **parallel do**
4: $\boldsymbol{w}_s \leftarrow$ OnlineLearn($\langle F_s, E_s \rangle$)
あるいはBatchLearn($\langle F_s, E_s \rangle$)
5: **end for**
6: $\boldsymbol{w} \leftarrow \mathrm{mix}(\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_S\})$
7: **return** $\boldsymbol{w}$
8: **end procedure**

shardへ分割

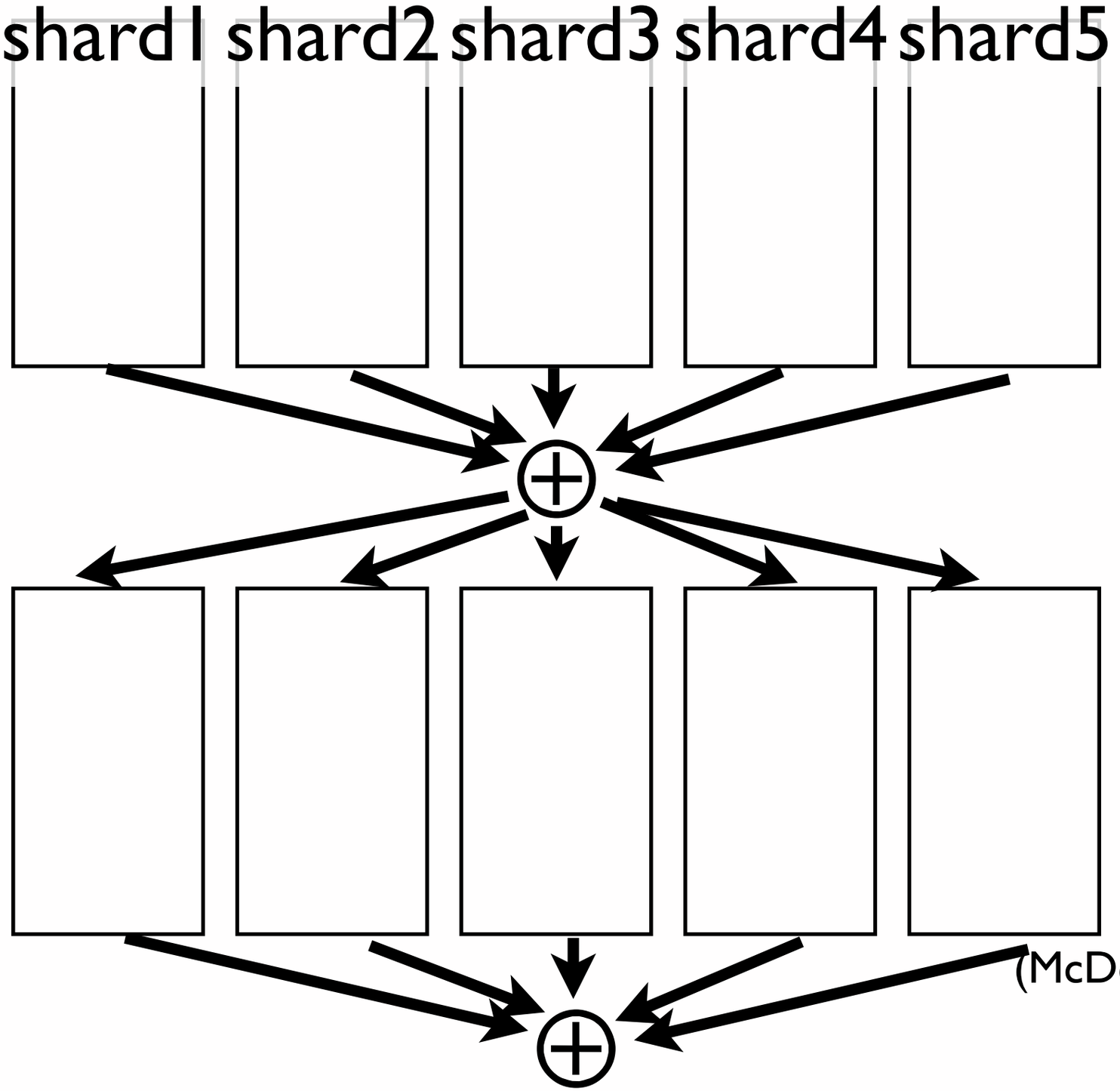
各shardで学習

混合

- 各shard独立に学習されるため、全体で最適解が保証されない

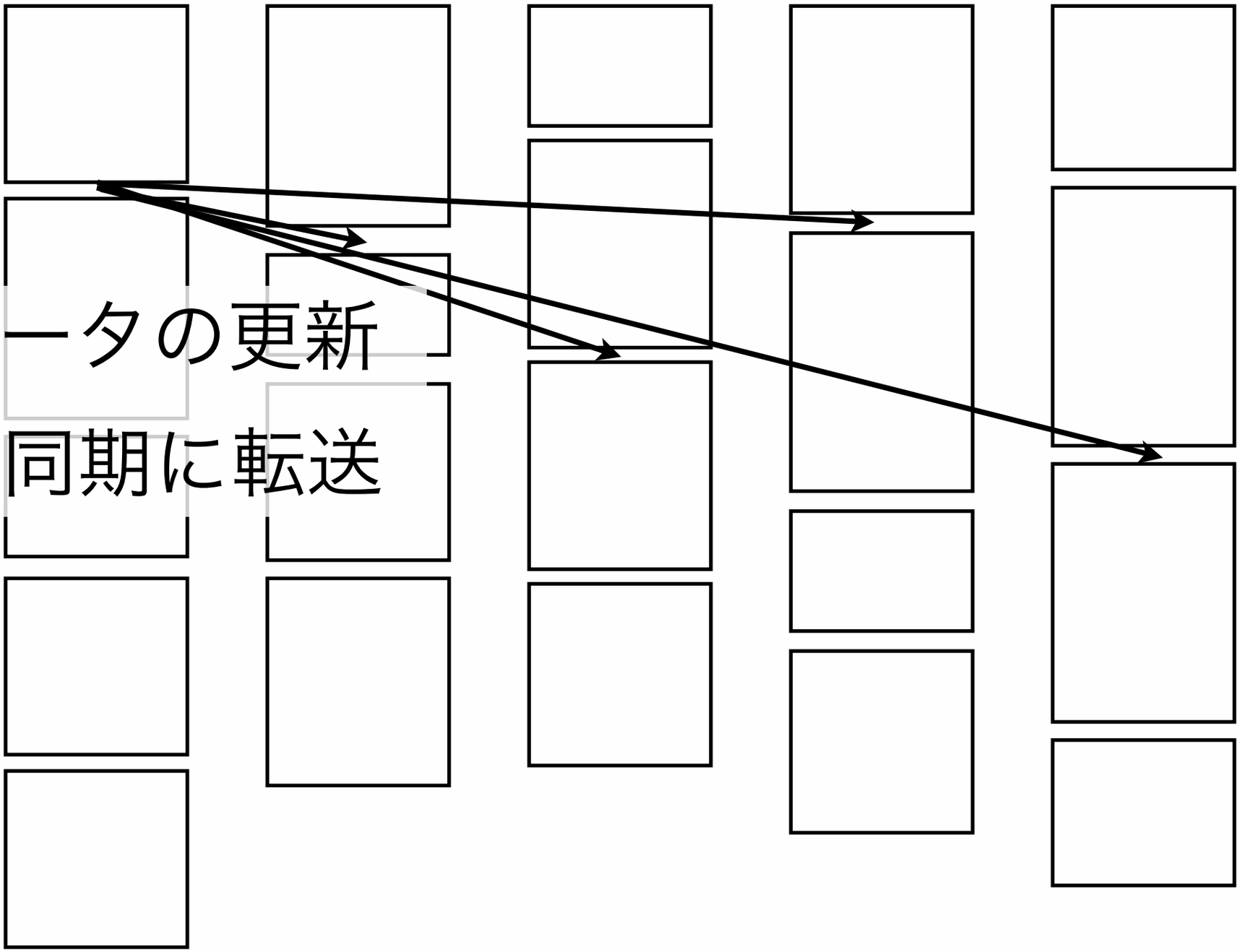
同期更新

繰り返し毎に混合



(McDonald et al., 2010)

非同期更新



パラメータの更新分を非同期に転送

まとめ

- オンライン学習: 大規模データ、高次元素性
- パーセプトロン、MIRA、SGD、AdaGrad
- 並列化による効率化

参考文献

- David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP 2008*, pages 224-233, Honolulu Hawaii, October.
- David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. of NAACL-HLT 2009*, pages 218-226, Boulder, Colorado, June
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceed ings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176-181, Portland, Oregon, USA, June.
- Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551-585 March.
- John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121-2159, July.

参考文献

- Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the As sociation for Computational Linguistics: Human Language Technologies*, pages 221-231, Montréal, Canada, June.
- Ryan McDonald, Keith Hall, and Gideon Mann. 2010. Distributed training strategies for the structured perceptron In *Proc. of NAACL-HLT 2010*, pages 456-464, Los Angeles California, June.
- Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistica machine translation. In *Proc. of ACL 2002*, pages 295-302 Philadelphia, Pennsylvania, USA, July.
- Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hidek Isozaki. 2007. Online Large-Margin Training for Statistica Machine Translation. In *Proc. of EMNLP-CoNLL 2007*, pages 764-773, Prague, Czech Republic, June.