Learning Fast-Mixing Models for Structured Prediction

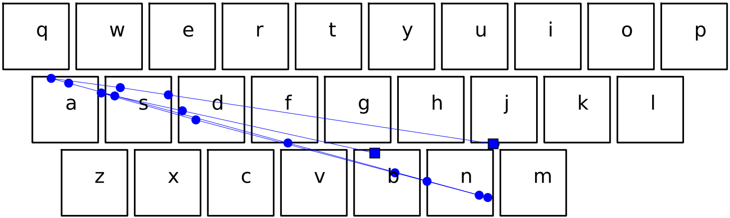
Jacob Steinhardt   Percy Liang

Stanford University

{jsteinhardt,pliang}@cs.stanford.edu

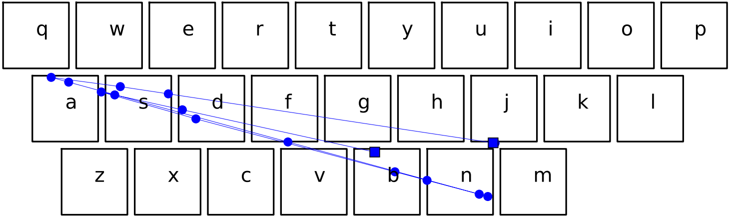
July 8, 2015

Structured Prediction Task




```
x:   b d s a d b n n n f a a s s j j j
z:   b # # a # # n-n-n # a-a # # n-n a
```

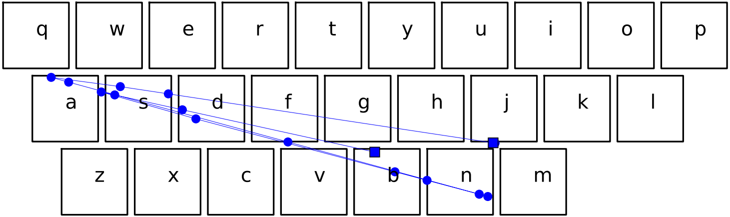
Structured Prediction Task




```
x:  b d s a d b n n n f a a s s j j j
z:  b # # a # # n-n-n # a-a # # n-n a
```

Goal: fit maximum likelihood model $p_\theta(z \mid x)$. Two routes:

Structured Prediction Task

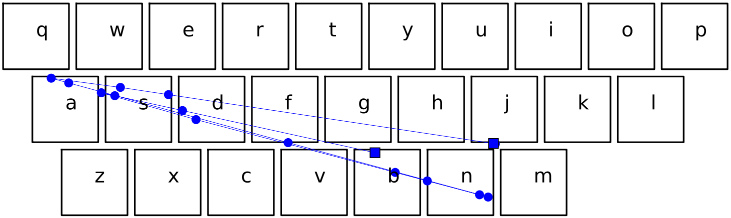



```
x:  b d s a d b n n n f a a s s j j j
z:  b # # a # # n-n-n # a-a # # n-n a
```

Goal: fit maximum likelihood model $p_\theta(z \mid x)$. Two routes:

- Use simple model u, exact inference

Structured Prediction Task



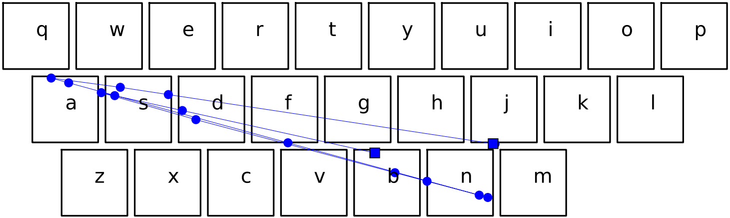
`x:  b d s a d b n n n f a a s s j j j`

`z:  b # # a # # n-n-n # a-a # # n-n a`

Goal: fit maximum likelihood model $p_\theta(z \mid x)$. Two routes:

- Use simple model u, exact inference
- Use expressive model, Gibbs sampling (transition kernel A)

Structured Prediction Task




```
x:  b d s a d b n n n f a a s s j j j
z:  b # # a # # n-n-n # a-a # # n-n a
```

Goal: fit maximum likelihood model $p_\theta(z \mid x)$. Two routes:

- Use simple model u, exact inference
- Use expressive model, Gibbs sampling (transition kernel A)

Can we get the best of both worlds?

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

u

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon) A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon) A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A \qquad u$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon) A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A \qquad u \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A \qquad u \longrightarrow A \longrightarrow A$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A \qquad u \longrightarrow A \longrightarrow A \qquad \cdots$$

Strong Doeblin Chains

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \longrightarrow A \longrightarrow A \longrightarrow A \qquad u \longrightarrow A \qquad u \longrightarrow A \longrightarrow A \qquad \cdots$$

All Doeblin chains mix quickly:

Proposition

If $\tilde{A}$ is ε strong Doeblin, then its mixing time is at most $\frac{1}{\varepsilon}$.

Definition (Doeblin, 1940)

A chain $\tilde{A}$ is *strong Doeblin* with parameter ε if

$$\tilde{A}(z_t \mid z_{t-1}) = \varepsilon u(z_t) + (1-\varepsilon)A(z_t \mid z_{t-1})$$

for some u, A.

$$u \to A \to A \to A \qquad u \to A \qquad u \to A \to A \qquad \cdots$$

All Doeblin chains mix quickly:

Proposition

If $\tilde{A}$ is ε strong Doeblin, then its mixing time is at most $\frac{1}{\varepsilon}$.

Moreover, the stationary distribution is $A^T u$, where $T \sim \text{Geometric}(\varepsilon)$.

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\tilde{A}_\theta = \varepsilon u_\theta + (1 - \varepsilon) A_\theta$$

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon)\underbrace{A_\theta}_{\pi_\theta}$$

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\begin{array}{ccc} \tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon) A_\theta \\ \tilde{\pi}_\theta \longleftarrow \pi_\theta \end{array}$$

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

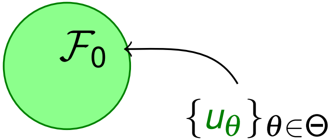
$$\tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon) A_\theta$$
$$\tilde{\pi}_\theta \longleftarrow \pi_\theta$$

Three model families:

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\begin{array}{ccc} \tilde{A}_\theta & = \varepsilon u_\theta + (1-\varepsilon) & A_\theta \\ \tilde{\pi}_\theta & \longleftarrow & \pi_\theta \end{array}$$

Three model families:

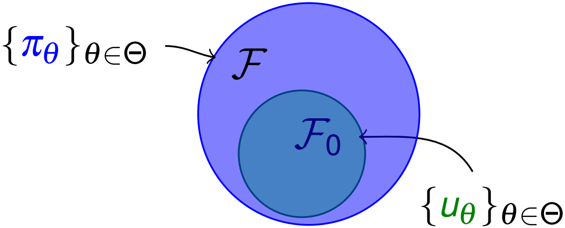


A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon) A_\theta$$
$$\tilde{\pi}_\theta \longleftarrow \pi_\theta$$

Three model families:

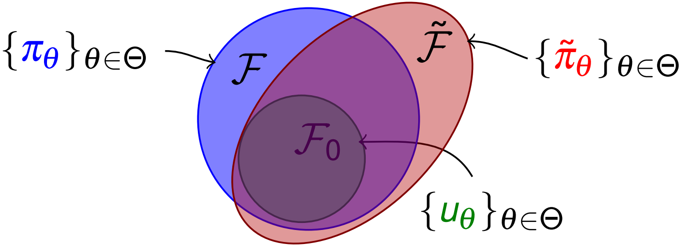


A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon) A_\theta$$
$$\tilde{\pi}_\theta \longleftarrow \pi_\theta$$

Three model families:

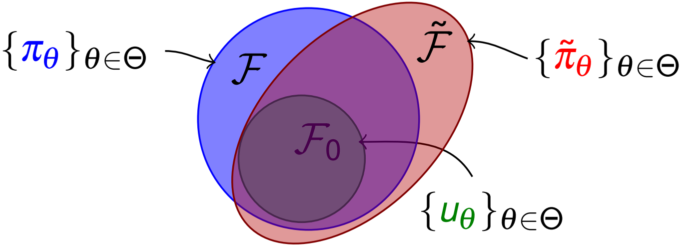


A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ.

$$\tilde{A}_\theta = \varepsilon u_\theta + (1-\varepsilon) A_\theta$$
$$\tilde{\pi}_\theta \longleftarrow \pi_\theta$$

Three model families:



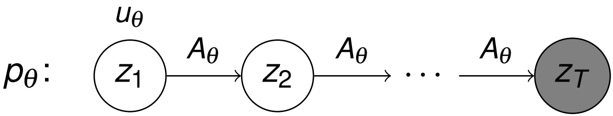
$\tilde{\mathcal{F}}$ **parameterizes computationally tractable distributions!**

Strategy

- Parameterize strong Doeblin distributions $\tilde{\pi}_\theta$
- Maximize log-likelihood: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \tilde{\pi}_\theta(z^{(i)})$
- Issue: hard to compute $\nabla L(\theta)$

Strategy

- Parameterize strong Doeblin distributions $\tilde{\pi}_\theta$
- Maximize log-likelihood: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \tilde{\pi}_\theta(z^{(i)})$
- Issue: hard to compute $\nabla L(\theta)$
- Insight: interpret Markov chain as latent variable model:



Strategy

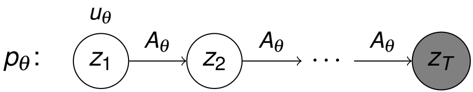
- Parameterize strong Doeblin distributions $\tilde{\pi}_\theta$
- Maximize log-likelihood: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \tilde{\pi}_\theta(z^{(i)})$
- Issue: hard to compute $\nabla L(\theta)$
- Insight: interpret Markov chain as latent variable model:

$$p_\theta : \quad \overset{u_\theta}{z_1} \xrightarrow{A_\theta} z_2 \xrightarrow{A_\theta} \cdots \xrightarrow{A_\theta} z_T$$

Observe: $\tilde{\pi}_\theta(z) = p_\theta(z_T = z)$, $T \sim \text{Geometric}(\varepsilon)$

Learning Updates

Recall latent variable model:



Learning Updates

Recall latent variable model:

$$p_\theta: \quad \overset{u_\theta}{z_1} \xrightarrow{A_\theta} z_2 \xrightarrow{A_\theta} \cdots \xrightarrow{A_\theta} z_T$$

Lemma

For any fixed z,

$$\frac{\partial \log p_\theta(z_T = z)}{\partial \theta} = \mathbb{E}_{z_{1:T-1} \sim p_\theta(\cdot \mid z_T = z)}\left[\frac{\partial \log p_\theta(z_{1:T})}{\partial \theta}\right].$$

Learning Updates

Recall latent variable model:

$$p_\theta: \quad \overset{u_\theta}{z_1} \xrightarrow{A_\theta} z_2 \xrightarrow{A_\theta} \cdots \xrightarrow{A_\theta} z_T$$

Lemma

For any fixed z,

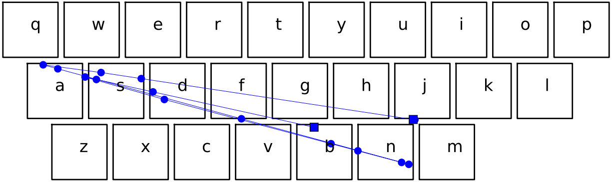
$$\frac{\partial \log p_\theta(z_T = z)}{\partial \theta} = \mathbb{E}_{z_{1:T-1} \sim p_\theta(\cdot \mid z_T = z)}\left[\frac{\partial \log p_\theta(z_{1:T})}{\partial \theta}\right].$$

Upshot: just need to sample trajectories that end at z.

$\Longrightarrow$ importance sampling

Experiments: Task

Task from before:

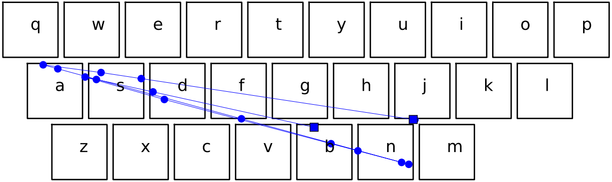



```
x:  b d s a d b n n n f a a s s j j j
z:  b # # a # # n-n-n # a-a # # n-n a
y:  b     a     n         a       n   a
```

Note y is a deterministic function $y = f(z)$.

Experiments: Task

Task from before:



x:	b	d	s	a	d	b	n	n	n	f	a	a	s	s	j	j	j
z:	b	#	#	a	#	#	n-	n-	n	#	a-	a	#	#	n-	n	a
y:	b			a			n				a				n		a

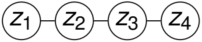
Note y is a deterministic function $y = f(z)$.

Goal: learn model $p(z \mid x)$ that maximizes

$$p(y \mid x) = \sum_{z \in f^{-1}(y)} p(z \mid x)$$

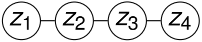
Experiments: Setup

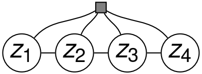
Models:

$u(z \mid x)$ (bigram, DP): 

Experiments: Setup

Models:

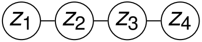
$u(z \mid x)$ (bigram, DP): 

$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs): 

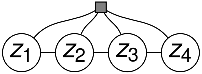
Experiments: Setup

Models:

$u(z \mid x)$ (bigram, DP):



$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs):

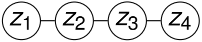


Comparisons:

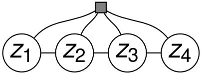
Experiments: Setup

Models:

$u(z \mid x)$ (bigram, DP):



$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs):



Comparisons:

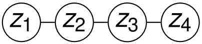
- basic: Gibbs sampling (A)

  (compute gradients assuming exact inference)

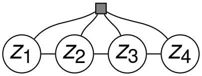
Experiments: Setup

Models:

$u(z \mid x)$ (bigram, DP):



$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs):



Comparisons:

- basic: Gibbs sampling (A)

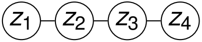
  (compute gradients assuming exact inference)

- u_θ-Gibbs: Gibbs with random restarts from u

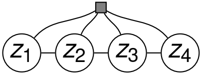
Experiments: Setup

Models:

$u(z \mid x)$ (bigram, DP):



$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs):

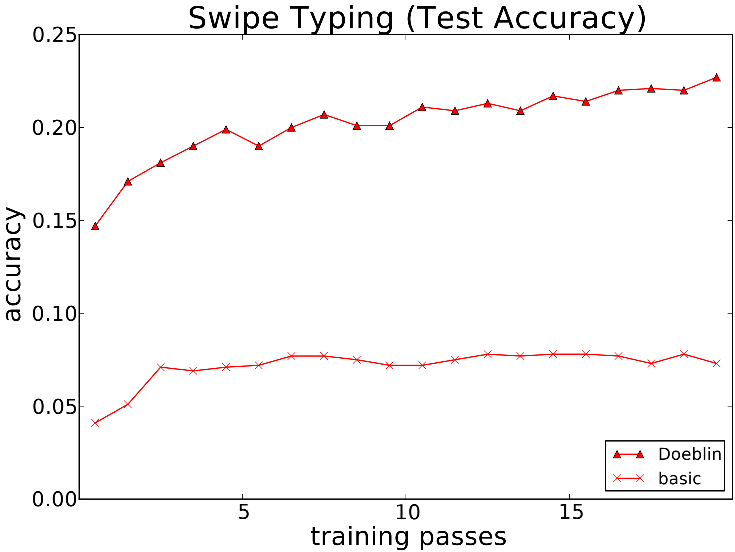


Comparisons:

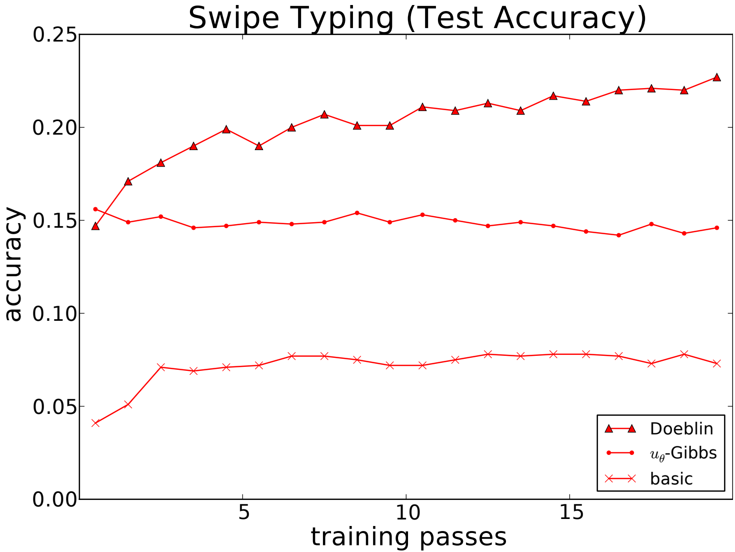
- basic: Gibbs sampling (A)

  (compute gradients assuming exact inference)
- u_θ-Gibbs: Gibbs with random restarts from u
- Doeblin: our method

Experiments: Results



Experiments: Results



Discussion

Summary:

- Strong Doeblin property enables fast mixing
- Interpolates between tractability and expressivity
- Provides better learning updates at training time

Discussion

Summary:

- Strong Doeblin property enables fast mixing
- Interpolates between tractability and expressivity
- Provides better learning updates at training time

Also in paper:

- Theoretical analysis of strong Doeblin family
- Multi-stage Doeblin chains

Discussion

Related work:

- Policy gradient (Sutton et al., 1999)
- Inference-aware learning (Barbu, 2009; Domke, 2011; Stoyanov et al., 2011; Huang et al., 2012)
- Strong Doeblin analysis (Doeblin, 1940; Propp & Wilson, 1996; Corcoran & Tweedie, 1998)

Discussion

Related work:

- Policy gradient (Sutton et al., 1999)
- Inference-aware learning (Barbu, 2009; Domke, 2011; Stoyanov et al., 2011; Huang et al., 2012)
- Strong Doeblin analysis (Doeblin, 1940; Propp & Wilson, 1996; Corcoran & Tweedie, 1998)

Future work:

- Explore other tractable families
- Learn multi-stage chains

Discussion

Related work:

- Policy gradient (Sutton et al., 1999)
- Inference-aware learning (Barbu, 2009; Domke, 2011; Stoyanov et al., 2011; Huang et al., 2012)
- Strong Doeblin analysis (Doeblin, 1940; Propp & Wilson, 1996; Corcoran & Tweedie, 1998)

Future work:

- Explore other tractable families
- Learn multi-stage chains

Reproducible experiments on CodaLab: `codalab.org/worksheets`