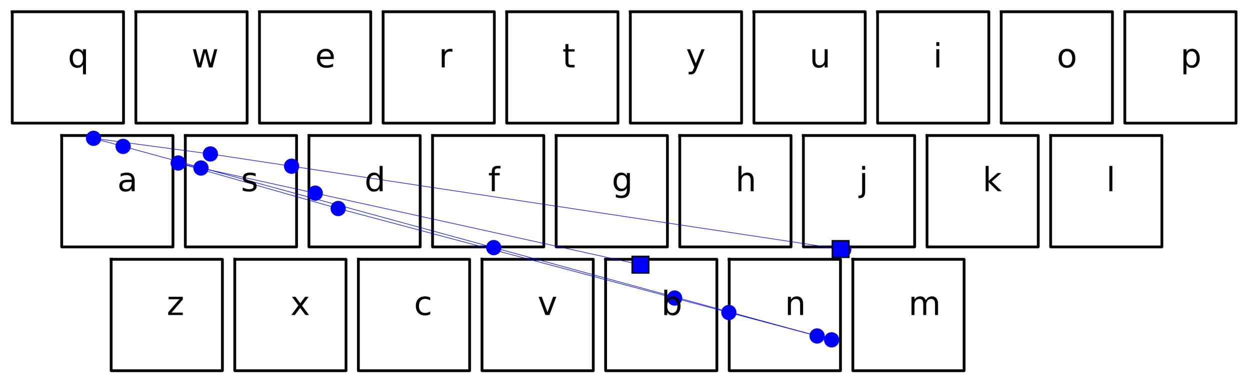
Learning Fast-Mixing Models for Structured Prediction

Jacob Steinhardt **Percy Liang**

{jsteinhardt,pliang}@cs.stanford.edu

Structured Prediction Task

Swipe typing:




```
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
```

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
>
> $$\tilde{A}(y_t \mid y_{t-1}) = \epsilon u(y_t) + (1-\epsilon)A(y_t \mid y_{t-1})$$
>
> 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 $\frac{1}{\epsilon}$.

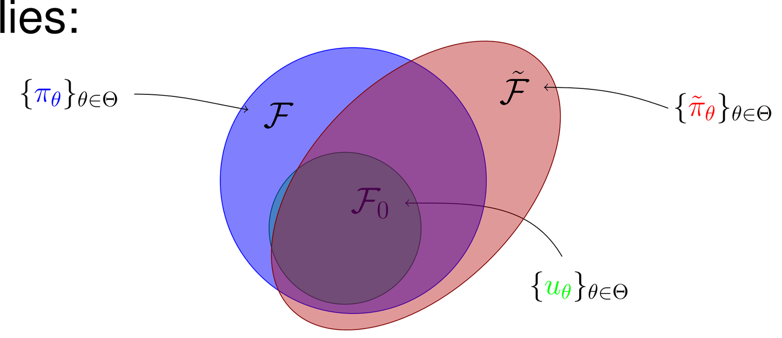
Moreover, the stationary distribution is $\mathbb{E}[A^T u]$, where $T \sim \mathrm{Geometric}(\epsilon)$.

A Strong Doeblin Family

Let θ parameterize a distribution u_θ and transition matrix A_θ. Define:

$\tilde{A}_\theta = \epsilon u_\theta + (1-\epsilon)A_\theta$	(strong Doeblin chain)
$\tilde{\pi}_\theta =$ stationary distribution of $\tilde{A}_\theta$	(tractable)
$\pi_\theta =$ stationary distribution of A_θ	(intractable)

Three model families:



Analysis of $\tilde{\mathcal{F}}$

How does $\tilde{\mathcal{F}}$ relate to $\mathcal{F}$?

First result: $\tilde{\mathcal{F}}$ approaches $\mathcal{F}$ as $\epsilon \to 0$.

> **Lemma**. For any fixed u, A, as $\epsilon \to 0$, $\mathrm{KL}\left(\tilde{\pi}_\theta \,\|\, \pi_\theta\right)$ and $\mathrm{KL}\left(\pi_\theta \,\|\, \tilde{\pi}_\theta\right)$ both approach 0 monotonically.

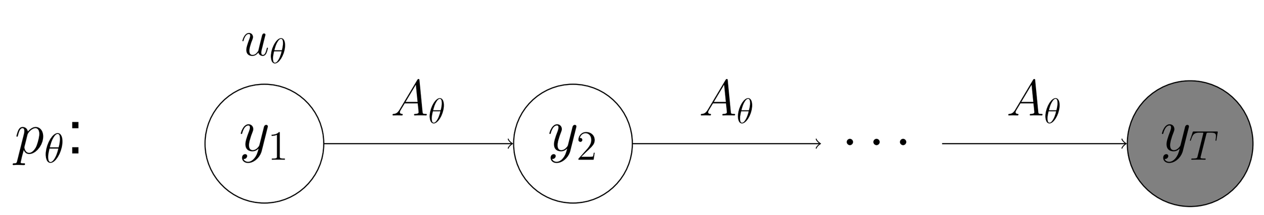
Second result: $\tilde{\mathcal{F}}$ well-approximates the elements of $\mathcal{F}$ with mixing time $\ll \frac{1}{\epsilon}$.

> **Lemma**. Let D_{χ^2} denote χ^2-divergence and $\gamma(A)$ be the spectral gap of A. Let π be the stationary distribution of A and $\tilde{\pi}$ be the stationary distribution of $\tilde{A}$. Then
>
> $$D_{\chi^2}(\pi \| \tilde{\pi}) \le \frac{\epsilon}{\gamma(A)} D_{\chi^2}(\pi \| u).$$

Also note: if each u has an A that leaves it invariant, then $\tilde{\mathcal{F}}$ contains $\mathcal{F}_0$. More generally, $\tilde{\mathcal{F}} \supseteq \mathcal{F} \cap \mathcal{F}_0$.

Maximum-Likelihood Learning

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



Observe: $\tilde{\pi}_\theta(y) = p_\theta(y_T = y)$, $T \sim \mathrm{Geometric}(\epsilon)$

Can now use standard lemma about marginal likelihood:

> **Lemma**. For any fixed y,
>
> $$\frac{\partial \log p_\theta(y_T = y)}{\partial \theta} = \mathbb{E}_{y_{1:T}\sim p_\theta}\left[\frac{\partial \log p_\theta(y_{1:T})}{\partial \theta}\middle| y_T = y\right].$$

Upshot: just need to sample trajectories that end at y.
$\Longrightarrow$ importance sampling

- sample $y_{1:T-1}$ unconditionally, importance weight by $A(y_T = y \mid y_{T-1})$
- can also assign weights to each prefix $y_{1:t-1}$ to reduce variance

Experiments

Structured prediction task from before (swipe typing, see first panel).
Note y is a deterministic function $y = f(z)$.

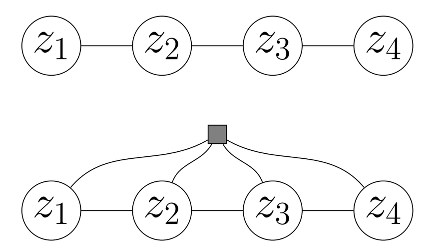
Goal: learn model $p(z \mid x)$ that maximizes

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

Models:

$u(z \mid x)$ (bigram, dynamic program):

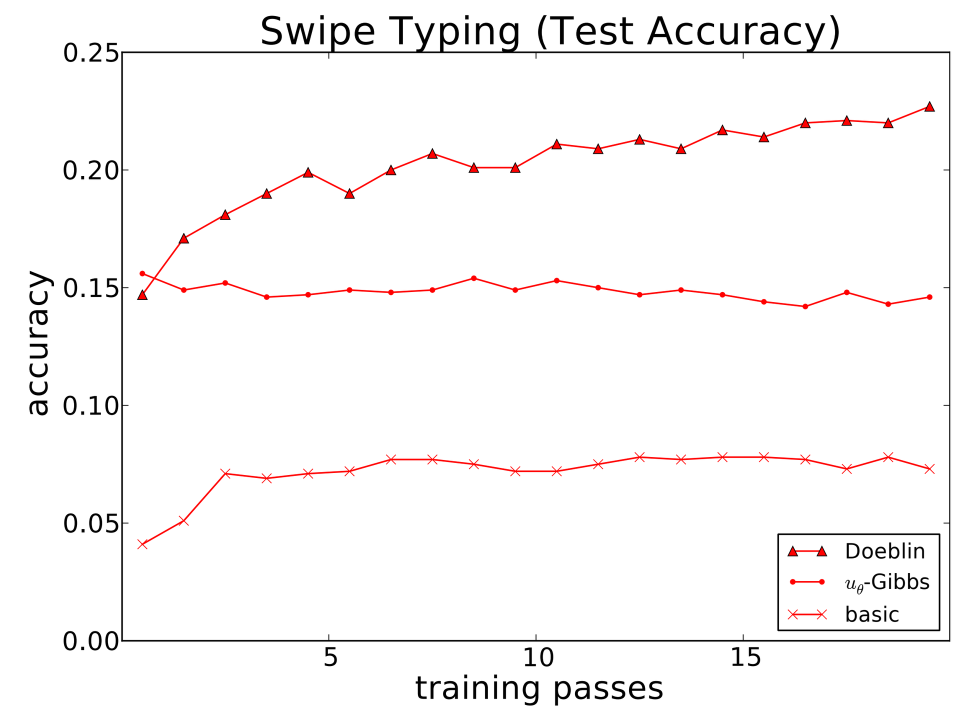
$A(z_t \mid z_{t-1}, x)$ (dictionary, Gibbs):



Comparisons:

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

Results:



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)

Reproducibility

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

The first author was supported by the Hertz foundation and by the NSF.