Sliding Windows — This is new in Yih's article

Matrix Sketching — Not that, that is old

$\varepsilon = \|A^TA - B^TB\|_2 / \|A\|_F^2$ space complexity $O(\frac{d}{\varepsilon})$

Task: compress $A \in \mathbb{R}^{N\times d}$ to $B \in \mathbb{R}^{l\times d}$ where $l \ll N$

What's new:

need extra assumption, $A_{i(\cdot)}$ is normalized

when A updates like a sliding window,

can we also update B to track the new A.

Frequent Directions:

Initially, B are all zero.

If we have zero rows, then just put a into B

If not, we first append w to B, and then apply SVD to shrink matrix B.

$$B = \begin{pmatrix} \sigma_1 - \sigma_l & & \\ & \ddots & \\ & & \sigma_{l-1} - \sigma_l \end{pmatrix} V^T_{[1..l-1]}$$

remove one row from V

Why don't they just compute A^TA and run SVD afterwards?

$A \in \mathbb{R}^{N \times d}$ $\qquad A^TA$

Useless shit $\times$

Restate the problem (Vector codebook)

Given a function $\underset{\mathbb{R}^{N\times d}}{\underset{\rotatebox{90}{$\in$}}{f(v)}} : \mathbb{R}^d \to \mathbb{R}^d$ (actually: $\{N : \mathbb{N}^+\} \to \mathbb{R}^{N\times d} \longrightarrow (\mathbb{R}^d \to \mathbb{R}^d)$)

Goal: find another function $\underset{\mathbb{R}^{o(N)\times d}}{\underset{\rotatebox{90}{$\in$}}{\hat{f}(w)}} : \mathbb{R}^d \to \mathbb{R}^d$

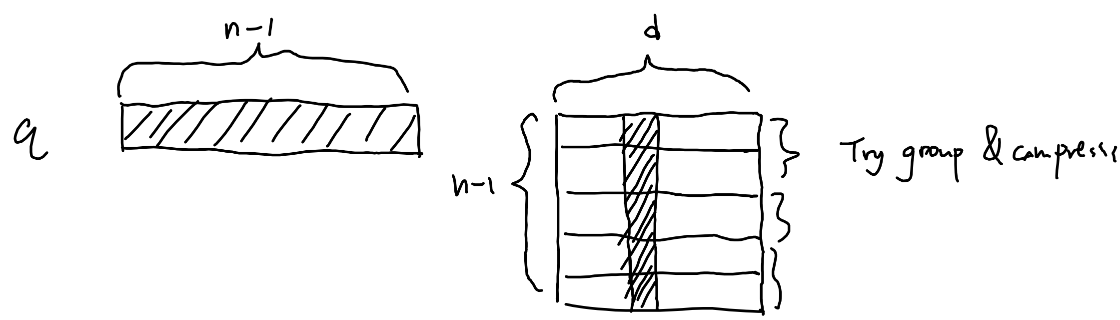
Expect some $p(N) = o(N)$ here

Heuristic: grab a machine learning model, update the model with v_n

$g(w + v_n) \to w$

such that $\hat{f}(w)$ behaves like $f(v)$

because $\hat{f}(w)$ approximates $f(v_{[1\ldots n-1]})$



$$f(w,q) = \mathrm{softmax}(q\, v^T_{[1\ldots n-1][1\ldots d]})\, v_{[1\ldots n-1][1\ldots d]}$$

$$f(w',q) = \mathrm{softmax}(q\, v^T_{[1\ldots n][1\ldots d]})\, v_{[1\ldots n][1\ldots d]}$$

$$f(w,q) \propto \exp(q\, v^T_{[1\ldots n-1][1\ldots d]})\, v_{[1\ldots n-1][1\ldots d]}$$

$$f(w',q) \propto \exp(q\, v^T_{[1\ldots n][1\ldots d]})\, v_{[1\ldots n][1\ldots d]}$$

Turn this into several constraints

$$\boxed{f_V(q) = \text{sofmax}(q V^T) V} \qquad V \in \mathbb{R}^{n \times d} \quad q \in \mathbb{R}^d$$

Don't view that as a direct computation problem, instead find a way to opeimize it as some differential equation.

(Ask Jan for some ideas, or juse read his paper)

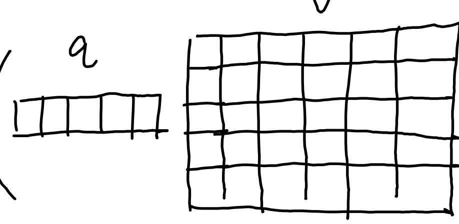
$v \to \phi(v)$ } embed v into some S

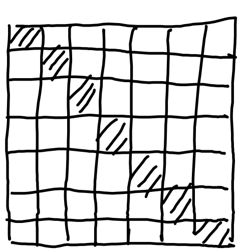
$\circ$ } a simple operation on S

$g(\cdot, \cdot)$ } a mapping from $\mathbb{R}^d \times S$ to $\mathbb{R}^d$
can have some iterations in it

Doesn't feel very correct ...

$$\text{softmax}(q V^T) = \frac{1}{Z} \cdot \exp\Big(\quad q \quad V^T \quad \Big)$$





$$= \exp(\text{diag}(q V^T)) \times \text{diag}(V_{\cdot j})$$

$$\underbrace{\exp(\text{diag}(q V^T))} \times V \times \mathbb{1}$$

$\exp(\underset{\sim}{A})$
$\uparrow$
$n \times n$

$$\frac{dx}{dt} = A\, x(t)$$

$t \;\; 0 \to 1$

Given q, v, we can run DE to obtain the result.

Fix the # of samples we use to compute the DE. ↵ Let it be m

We just get to compute a telescopic sum for each v_{ij}

$$g_t((qv^T)_i, v_{ij}) = g_{t-1}((qv^T)_i, v_{ij})\left(\frac{1}{m} + (qv^T)_i\right)$$ 不太有用

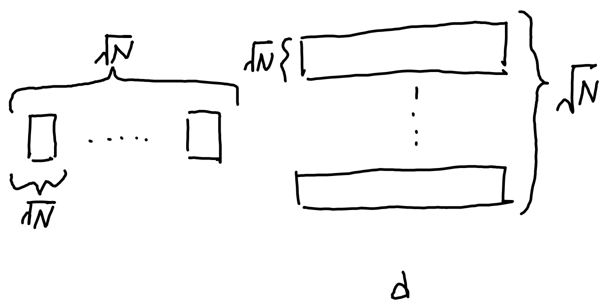
---

q is incoming v

不知道有啥用处

$$f(v, q) = \mathrm{softmax}(qKv^T)v$$

then, $v \leftarrow \begin{pmatrix} v \\ q \end{pmatrix}$



不知道引入什么误差能有什么收益

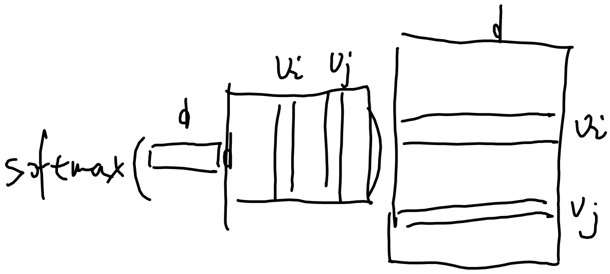
对新加入的v，找一个簇

然后找一个系数

Add more assumptions : v and q are both upperbounded by M

Simpler question : what happens when we merge v_i & v_j ?

what is optimal λ for i, j, v, q ?



$$\frac{\sum_l \exp(qv_l^T)v_l}{\sum_l \exp(qv_l^T)} = \frac{\sum_{l \setminus \{i,j\}} \exp(qv_l^T)v_l + v_i\exp(qv_i^T) + v_j\exp(qv_j^T)}{\sum_{l \setminus \{i,j\}} \exp(qv_l^T) + \exp(qv_i^T) + \exp(qv_j^T)}$$

softmax$(qKv^T)v$

q 和 softmax$(qv^T)v$ 有多像？

因为是算加权平均，而和 q 最接近的 v_i 分到最大的权重

$\hat{q} = qv^T(vv^T)^{\#}v$ ← N 不大的时候，我猜是这种

什么时候可以把 v_i 和 v_j 合并？ ~~需要~~ v_i 和 v_j 各维都差不多

分开考虑 $v_{:,d}$ 可能更加合适

If $v_{i,:}$ is normalized, i.e. $\|v_{i,:}\|_2 = 1$

$$(\text{softmax}(qv^T))_i = \frac{\exp\left(-\frac{\|q-v_{i,:}\|^2}{2}\right)}{\sum_j \exp\left(-\frac{\|q-v_{j,:}\|^2}{2}\right)}$$

For each dimension $d \in [1..D]$

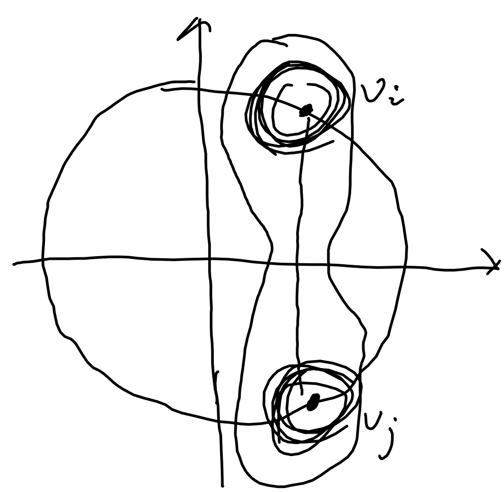
Compress or filter out Gaussians.

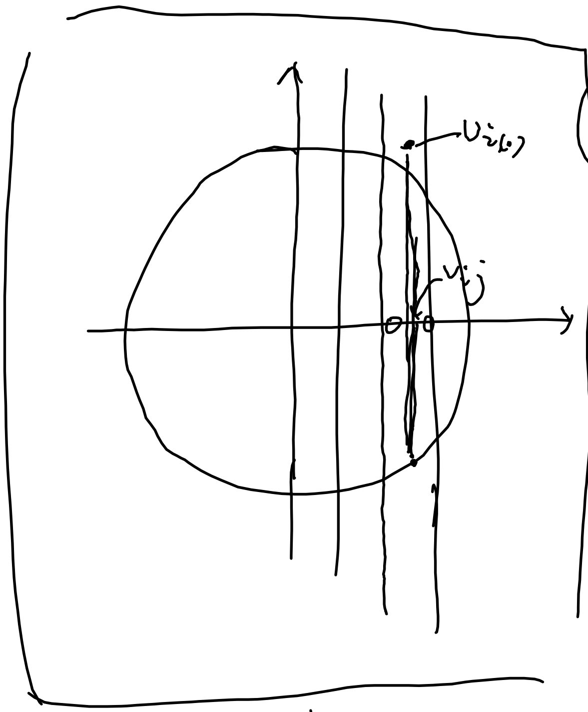
$v_{i1} = v_{j1}$

$$p_{I,X}(i, \underset{\sim}{x}) = \frac{1}{N}\exp\left(-\frac{\|x-v_i\|^2}{2}\right)$$

$i \in [1 \dots N]$

$\mathbb{E}[v_I \mid X = q]$





$\times d$

$$f_v(q)_j = \left(\quad \right) \cdot v_{(\cdot),j} \quad \times d$$

$j \in [1 \ldots d]$

$$\int_{x \in \mathbb{R}^d} \exp\left(\frac{x^T x}{2}\right) dx$$
$$= \int_{r \in \mathbb{R}^+} \exp\left(\frac{r^2}{2}\right) f(r)\, dr^2$$

p.d.f. (no degen…)

$$f_1 \searrow p(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma_p)}} \exp\left(-\frac{(x-\mu_p)^T \Sigma_p^{-1} (x-\mu_p)}{2}\right)$$

$$f_0 \searrow q(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma_q)}} \exp\left(-\frac{(x-\mu_q)^T \Sigma_q^{-1} (x-\mu_q)}{2}\right)$$

$(X, I) \sim (f_i, i)$

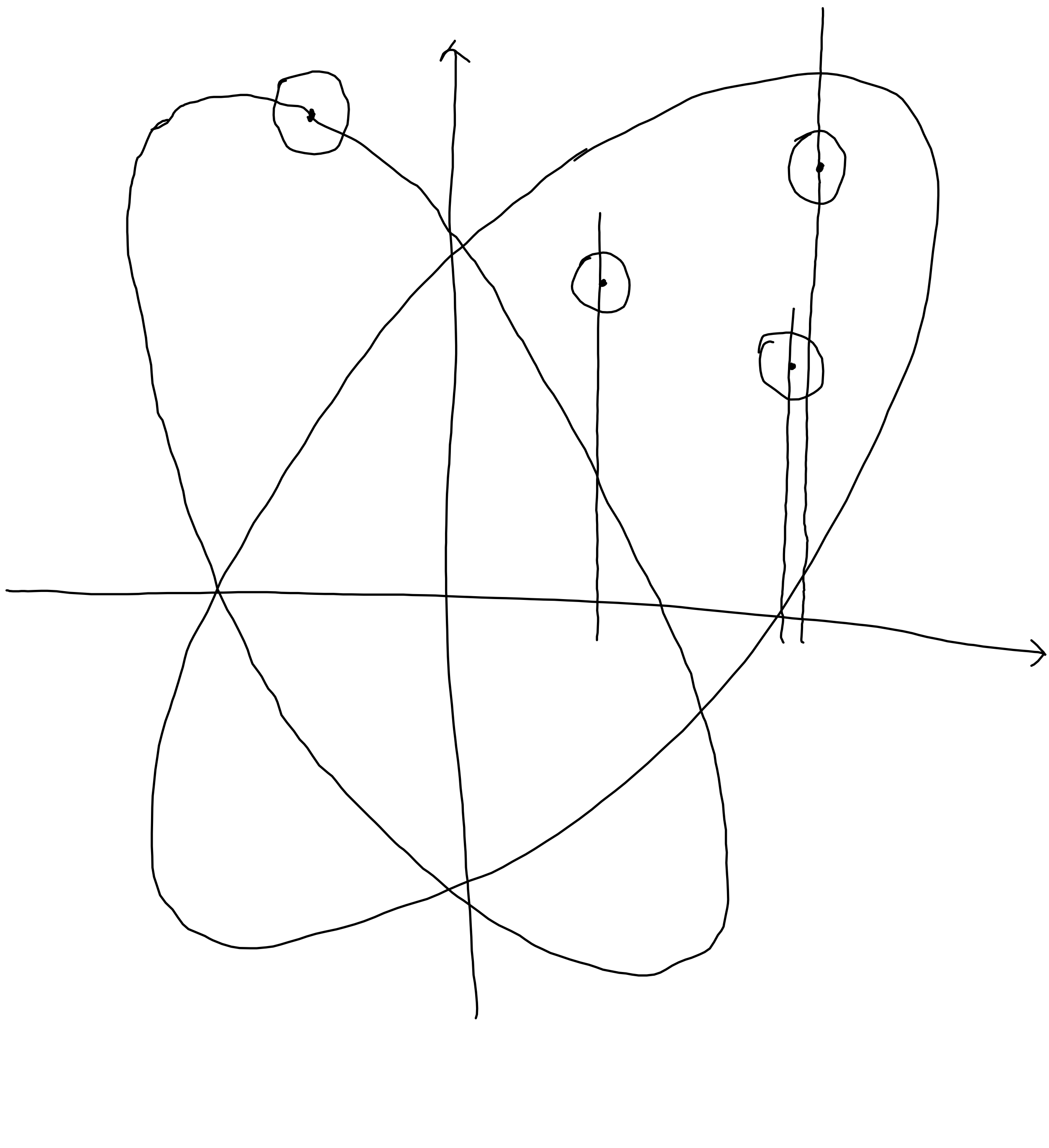
$\mathbb{V}$ for Variance

$\|\mu_p - \mu_q\|_2$

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X \mid I]]$$
$$= \frac{1}{2}\mathbb{V}[X \mid I = 1] + \frac{1}{2}\mathbb{V}[X \mid I = 0]$$
$$= \frac{1}{2}(\Sigma_p + \Sigma_q)$$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid I]]$$
$$= \frac{1}{2}(\mu_p + \mu_q)$$

$$\sum_{(k,v)\in C} \exp(-q^Tq/2 + q^Tk - k^Tk/2) \cdot (v \cdot \exp(k^Tk/2))$$



MCMC

Example : estimate π

$$X, Y \sim \text{Uniform}(-1, 1)$$

$$Z = [X^2 + Y^2 \le 1]$$

$$\Pr\{Z = 1\} = \frac{\pi}{4}$$

Theorem : Chernoff Bound

$\{X_1, \cdots, X_m\}$ $\quad \mu = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[X_i]$

$X_i(\omega) \in \{0, 1\}$ $\quad \Pr\left\{ \left| \frac{1}{m} \sum_{i=1}^{m} X_i - \mu \right| \ge \varepsilon \mu \right\} \le \delta \Leftarrow m \ge 3(\ln(2/\delta))/\varepsilon^2 \mu$

Definition : FPRAS (ε, δ) $\quad \varepsilon \quad \ln \delta^{-1}$ size of input

Example : DNF Counting $\quad f(x_1, \cdots, x_n) = (x_1 \wedge \neg x_2 \cdots \wedge x_n) \vee (\cdots\cdots)$

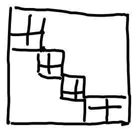
disjunction (or) $\quad \{0, 1\}$ $\quad$ "x_i" or "$\neg x_i$" joined by "$\wedge$"

$$c(f) = \{(x_1, \cdots, x_n) : f(x_1, \cdots, x_n) = 1\}$$

$$\mathbb{E}[f(X_1, \cdots, X_n)] = \frac{c(f)}{2^n}$$

$$X_i \sim \text{Uniform}(0, 1)$$

To give an (ε, δ)-estimation of $c(f)/2^n$, samples $m \ge \dfrac{3 \cdot 2^n \cdot \ln(2/\delta)}{\varepsilon^2 c(f)}$

$$\exp\left(q\,Q\right.$$  $$\left.K^{T} x^{T}\right) x V$$

The type theory for notiz language

① I need the evaluation to be very efficient.

i.e. For a function term $x \to \underline{\underline{y}}$. $\underline{\underline{\text{This}}}$ syntax tree is traversed only once for subst x.

which means no reduce happen unless y has an "endpoint" type.

To keep track of values, we use a "Stuck Variant" to store the value and removed abstraction layers.

This is a bad idea. Because we have tuples.

② $\underline{\underline{\text{CPS}}}$ to rescue

Continuation Pass Style

Idea: for each term, find a way to represent "the next step"

Goal: for each term M, convert it to a term $CPS[M]$, such that $CPS[M, k]$ means when k is eventually called its argument $= M$.

$$CPS[(M\ N), k] := CPS[N, CPS[M, k]]$$

$$CPS[\lambda x. M, k] := \lambda x. CPS[M, k]$$

$$CPS[N, \lambda x. CPS[M, k]] \equiv CPS[M[x/N], k]$$

This will cause an ever-growing stack.

So no X

③ De Brujin with <u>move</u>

Define "last usage" of a variable, use movement

Case 1:

$\lambda x.(\underset{\substack{?\\ \uparrow \\ x}}{M}\ \underset{\substack{\sim \\ \uparrow \\ x}}{N})$ then "last usage" is in N

(M is evaluated first)

Case 2:

$\lambda x.(\underset{\substack{\sim \\ \uparrow \\ x}}{M}\ N)$ then "last usage" is in M

Case 3:

$\lambda x.(\lambda y.\underset{\substack{\sim \\ \uparrow \\ x}}{M})$ then "last usage" is in M

④ Composition

1. Modules are tree shaped (like rust)
2. Module-level parameters: different params results in different views when compiling standalone modules (code analysis).
3. Modules are imported and included with file path.
4. Module attributes are constructed using export statement (top level, "only once", same for mod-params)
5. "top level" is defined during lowering (before evaluation)

⑤ Defining Algebraic Data Types

Trouble: Consider evaluating function $f: A\times B \to C$

The problem is $\overset{\text{or } A}{B}$ is not necessarily binded in for example

$\lambda x.(f\ (a\ x))$, we shift x to stack,

but where to store a.

The evaluation should be designed against it.

Allow enum/tuple/struct types.

Markups will be encoded into these types.

Build core terms from face sytax.

During execution, values are somewhat byte encoded.

Divide terms to $\begin{cases} \text{values} \\ \text{neutrals} \\ \text{reducibles} \end{cases}$

⑥ Clarify the misunderstanding of CPS

1. During CPS evaluation, it does not produce neutral terms.
2. Not free access is passed to continuation, so no stack is involved.

⑦ Leave shallow captures and lift deep captures.

Neutrals should not be visited until they turn into values.

⑧ Modular design:

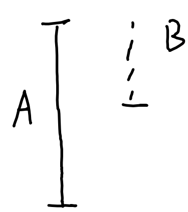
Each module is defined (applied) only once.

A module must be materialized before import.

In a file, dependency between modules can be determined by the appearance of import and module arguments.

If we import a module A before we define module B, then B is dependent on A. Graph maintains start/end of a module.

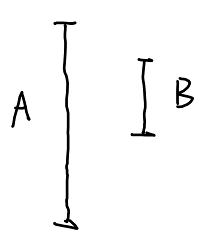
import (A; B) module A { import B; ... }



B must end before compile A ends.

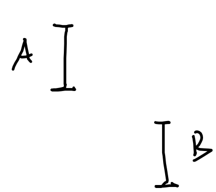
Good: module A ( );
module B ( );

define (A; B) module A { module B(...); ... }



B starts/ends within A

after (A; B) ∃ module C { import A; module B(...); }



A compiles before B starts

⑨ Compiler Behaviour on Module Updates

Sps. we have a dep graph G, when updating a module, we first remove the original module from that graph, and mark the modules defined in that module and down stream modules as dirty.

Then, we add modified modules back and generate edges in graph G'.

During that process, check each module s.t. they have one and only one definition, and no dependency loop occurs.

Finally, run each module w.r.t. G', update results when things are finished.

Dirty modules will not be re-evaluated if their input is equal.

⑩ Offloading: one observation is that exports are small but module output is large. We maintain an LRU-cached file system for that.

When evaluating include and import arguments, we always use the LRU-cached.

During evaluation, values are also offloaded to file system if they are big.

⑪ Module parameters on use v.s. on def

We may want module parameters to be different when we include modules.

* module parameters cannot appear in exported items

↑ This makes it so much easier

ref: reference to a codeblock in module

Allow generate state

⑫ How capturing works?

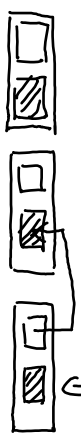
⑬ With pointer to stack top

i means %rsp $- i$

so the "inner" translates as usual,

but the "outer" is captured in outer variables

⑭ extract all variables in one pass



Because we assume every nested function escapes

Make sure everything is captured

During execution, this is everything we see.

When a symbol is discovered, we first find its definition level.

Then, we put the symbol along each level of current expressions' definition.

$[\quad]\lambda(x, y, z). \ldots$
$[\quad]\lambda(w, v, u). \ldots$
$(\ldots y \ldots)$

$\Downarrow$

$[\ldots]\lambda(x, y, z). \ldots$
$[y \ldots]\lambda(w, v, u). \ldots$
$(\ldots y \ldots)$

add it here

The captured variables should link to their immediate upper level.

Problem: if we have multiple levels for insertion, we may have to shift previous de Brujin indices.

Exact counting of Perfect Matching on Planar Graph

Theorem: Exactly counting PM in bipartite graph is #P-hard

Ex: #SAT / count $f: \{0,1\}^N \to \{0,1\}$

$$\{x : f(x) = 0\}$$

Definition: A Fully Poly-time Randomized Approx Scheme (FPRAS) for estimating Z, with ε, δ error:

$$\Pr\left((1-\varepsilon) Z \le \hat{Z} \le (1+\varepsilon) Z \right) \ge 1 - \delta$$

in time $\mathrm{poly}(n, \frac{1}{\varepsilon}, \log(\frac{1}{\delta}))$

Definition: A F P A Sampler for target distribution μ, a randomized algo, with ε error:

$d_{TV}(\nu, \mu) \le \varepsilon$ in $\mathrm{poly}(n, \log(1/\varepsilon))$

(Thm) $\exists$ FPRAS & FPAS for MP in bipartite graphs.

(lemma) FPRAS for $|\mathcal{M}|$ on all graphs ($\mathcal{M}$: matchings)

$\Leftrightarrow$ FPAS for μ on all graphs

$\downarrow$

uniform distribution over $\mathcal{M}$

Proof. $\frac{1}{|\overline{\mathcal{M}}|} = \Pr_{M\sim\mu}(M=\phi)$ $\quad E=\{e_1, \cdots, e_m\}$

($\Leftarrow$)

$= \Pr_{M\sim\mu}(e_1 \notin M \wedge \cdots \wedge e_m \notin M)$ decompose $\Pr(e_i \notin M | \cdots)$

$= \underbrace{\prod_{i=1}^{n} \Pr_{M\sim\mu/\{e_1,\cdots e_{i-1}\}}(e_i \in M)}$

estimate it with error ε/m

then est the count failure probablity δ/m

($\Rightarrow$) For $i=1$ to m: $\quad$ treat $M \in \overline{\mathcal{M}}$ as $\sigma \in \{0,1\}^m$

$p_i = \Pr_{\sigma\sim\mu}(\sigma_i = 1 \mid \sigma_1 \cdots \sigma_{i-1})$

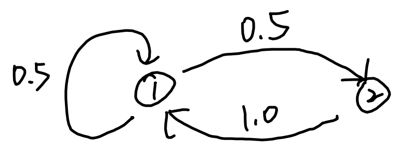
$= \frac{|\overline{\mathcal{M}}_{G' \setminus \{u_i, v_i\}}|}{|\overline{\mathcal{M}}_{G'}|}$ $\quad e_i = \{u_i, v_i\}$

Running tim will be $\mathrm{poly}(m, \frac{1}{\varepsilon})$

Boost to $\mathrm{poly}(m, \log(1/\varepsilon))$ with rejection sampling

(Def) Finite Markov Chain

(= random walk on directed graph)



(Def) A discrete-time & time-homogeneous Markov chain of a finite state space Ω is a sequence of R.V. $(X_t)_{t=0}^{\infty}$ valued in Ω satisfying:

$$\Pr(X_{t+1}=y \mid X_t=x \cdots X_0) = \Pr(X_{t+1}=y \mid X_t=x) = P(x,y)$$

Remark: row sum $\sum_y P(x,y) = 1$

Basic Properties:

① $\Pr(X_{t+2}=y \mid X_t=x)$

$$= \sum_z \Pr(X_{t+1}=z, X_{t+2}=y \mid X_t=x)$$

$$= \sum_z \Pr(X_{t+2}=y \mid X_{t+1}=z, X_t=x) \cdot \Pr(X_{t+1}=z \mid X_t=x)$$

② $\Pr(X_{t+l}=y \mid X_t=x) = P^l(x,y)$

③ if $X_0 \sim \mu_0$ then $\Pr(X_t=y) = (\mu_0 P^t)(y)$

(Def) stationary disribution of Markow Chain μ

(P : transition matrix)

$$\pi P = \pi$$

Every Markov chain has at least one stationary distribution.

($P\mathbb{1} = \mathbb{1}$; take the row eigenvector of P with eigenvalue 1;

Perron-Frobenus Theorem)

Prf: Take row vector $vP = v$

Consider $|v| := \begin{pmatrix} |v_1| \\ |v_2| \\ \vdots \\ |v_n| \end{pmatrix}$

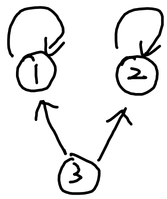
Triangle Ineq: $|v| = |vP| \leq |v|P$ $\forall$ entries

Also, $|v|P\mathbb{1} = |v|\mathbb{1}$

$\Rightarrow$ $|v| = |v|P$

Q: When can we find a unique stationary distribution?

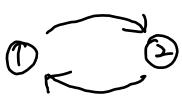
Bad examples:



The graph must be strongly connected.

$\forall x, y \in \Omega : \exists t \in \mathbb{N}$ s.t. $P^t(x,y) > 0$

Bad examples:



The graph must be aperiodic.

$\forall x \in \Omega$: period of $x := \gcd\{t \in \mathbb{N} : P^t(x,y) > 0\}$
$= 1$

$\forall x \in \Omega \quad p(x,x) > 0 \Rightarrow$ aperiodic

Therefore: $P' = \frac{1}{2}(I + P)$ is aperiodic.

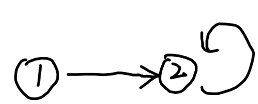
(Def) A finite MC is ergodic if it's aperiodic and irreducible.

$\Leftrightarrow \quad \exists t \in \mathbb{N}$ s.t. $\forall x, y \in \Omega : P^t(x,y) > 0$

Fundamental Theorem of Markov Chains:

$\forall$ ergodic finite MC P, $\exists$ unique $\pi P = \pi$

$\wedge \quad \lim_{t\to\infty} P^t = \mathbf{1}\pi = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}$



(Def) Total variation distance

$\forall$ distributions μ & π on finite Ω

$$d_{TV}(\mu, \pi) = \sum_{x \in \Omega} \mu(x) - \pi(x)$$

Property: $0 \leq d_{TV}(\mu, \pi) \leq 1 \ \wedge \ d_{TV}(\mu, \pi) \leq d_{TV}(\mu, \nu) + d_{TV}(\nu, \pi)$

(Def) A distribution ω is a coupling of (μ, π) s.t.

$$\forall x \in \Omega \quad \sum_{y \in \Omega} \omega(x, y) = \mu(x)$$

$$\forall y \in \Omega \quad \sum_{x \in \Omega} \omega(x, y) = \pi(y)$$

(Def) Identity coupling for μ and μ :

$$\omega(x, y) = \begin{cases} \mu(x) & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

(Def) Product coupling for μ and π :

$$\omega(x, y) = \mu(x) \cdot \pi(y)$$ independent case

(Lemma) Coupling lemma :

(a) $d_{TV}(\mu, \pi) \leq \sum_{x,y} \omega(x, y)$ if ω is a coupling of (μ, π)

(b) $\exists\, \omega$ s.t. it takes "="

(Proof of (b)) Let $\theta = \sum_{x} \min\{\mu(x), \pi(x)\}$

Set $\omega(x, x) = \min\{\mu(x), \pi(x)\}$

(Def) Consider a MC on Ω with transition matrix P

and two instances of (X_t, Y_t) ... joint distribution (X_t, Y_t)

(Proof) Fundamental theorem of MC (uniqness)

$\forall x, y \in \Omega$ Ergodic means $P^t(x, y)$

Consider two MC (X_t) & (Y_t)

$X_0 \sim \mu$ $Y_0 \sim \pi$

$\forall t \in \mathbb{N}$, $X_t \sim \mu$, $Y_t \sim \pi$

Construct a coupling of X_t, Y_t

If $X_{t-1} = Y_{t-1} = x$ then

Let $X_t \sim Q(x, \cdot)$ and $Y_t = X_t$

If $X_{t-1} = x \neq y = Y_{t-1}$

Let $(X_t, Y_t) \sim Q(x, \cdot) \otimes Q(y, \cdot)$

$Pr(X_t = Y_t \mid X_{t-1} = x, Y_{t-1} = y') \geq \varepsilon$

$d_{TV}(\mu, \pi) \leq Pr(X_t \neq Y_t)$

(Def) Pfaffian of a skew-symetric matrix

$$\mathrm{pf}(A) = \frac{1}{2^n n!} \sum_{\sigma \in S_{2n}} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(2i-1)\sigma(2i)}$$

$$\mathrm{pf}^2(A) = \det(A) \qquad \left(\qquad \right)$$

Motivation: to avoid sum over all possible permutations

find

(Lemma) If P is a symmetric matrix, then $\Omega = \text{uniform}(\Omega)$

(Proof) $P\mathbf{1} = \mathbf{1} \Rightarrow P^T\mathbf{1} = \mathbf{1}$

so $\mathbf{1}^T P = \mathbf{1}^T$, $(\frac{1}{|\Omega|}, \cdots \frac{1}{|\Omega|})$ is the stationary distribution

(Definition) P is reversible w.r.t. π if

$$\forall x, y \in \Omega : \pi(x) \cdot P(x,y) = \pi(y) \cdot P(y,x)$$

(Lemma) If P is reversible w.r.t. π then π is stationary

$$(\pi P)(x) = \sum_y \pi(y) P(y,x) = \sum_y \pi(x) P(x,y) = \pi(x)$$

(Example) Random walk on undirected graph $G = (V, E)$

$$P(u,v) = \begin{cases} \frac{1}{\deg(u)} & \text{if } \{v,u\} \in E \\ 0 & \text{otherwise} \end{cases}$$

Ergodic means connected & not bipartite in this case

$$\pi(u) = \frac{\deg(u)}{2|E|}$$

(Definition) For $\varepsilon \in (0,1)$, $T_{mix}(\varepsilon, P) = \max_{x \in \Omega} \min \{ t \in \mathbb{N} : d_{TV}(P^t(x, \cdot), \pi) \leq \varepsilon \}$

(Lemma) $T_{mix}(\varepsilon) \leq T_{mix}(\frac{1}{4}) \lceil \log_2(1/\varepsilon) \rceil$

(Example) Random walk on hypercube $H_n = \{0,1\}^n$

From $x_t \in \Omega$, uniformly pick a bit and mutate it with $p = \frac{1}{2}$

$$\forall x : P(x,x) \geq \frac{1}{2}$$

$P(x,y) = P(y,x)$ so it is symmetric

(Theorem) Mixing Time of hypercube H_n

Pick two chains with H_n distribution (X_t, Y_t)

Sps. at t, $X_t(w) = Y_t(w)$ If two bits are equal, they stay equal.

$$d_{TV}(P^t(x,\cdot), \pi) \le \Pr(X_t \ne Y_t)$$
$$\le \mathbb{E}[d_H(X_t, Y_t)]$$
$$\le (1-\tfrac{1}{n})^t n \le \tfrac{1}{4}$$

$$T_{mix}(\tfrac{1}{4}) = O(n \log n)$$

(Example) Proper Coloring

$G = (V, E)$ graph with max degree Δ

$\Omega \subseteq [q]^V$: a set of q-colorings of G, $q \ge \Delta$

s.t. $\forall \sigma \in \Omega$: $\sigma(v) \ne \sigma(u)$ for $uv \in E$

Goal : sample a q-coloring u.a.r. from G.

Choose $v \in V$ u.a.r. & $c \in [q] \setminus \underbrace{X_t(N(v))}_{\text{color of neighbors}}$

Set $X_{t+1}(w) = \begin{cases} c & \text{if } w = v \\ X_t(w) & \text{if } w \ne v \end{cases}$

(Lemma) If $q \ge \Delta + 2$, then GD is ergodic

& $\pi = \mathrm{unif}(\Omega)$ is stationary

(Theorem) If $q \ge 2\Delta + 1$ then $T_{mix} = O(n \log n)$

(Proof) Path Coupling

1. Construct $X_t = Z_0 - Z_1 - \cdots - Z_n = Y_t$

$$Z_i(v_j) = \begin{cases} Y_t(v_j) & \text{if } j \le i \\ X_t(v_j) & \text{if } j > i \end{cases}$$

2. Construct a coupling

3. "Glue"

(Theorem) Path Coupling

P : MC on finite Ω

Suppose $T = (\Omega, S)$ is a connected graph

$w : S \to [1, +\infty)$

$$d(x,y) = \min_{\substack{P:\text{path connecting } x \text{ and } y \\ \text{in } T}} \sum_{e \in P} w(e)$$

If $\forall x_0, y_0 \in S$, $\exists$ coupling of transition $(X_0, Y_0) \xrightarrow{P} (X, Y)$

s.t. $\mathbb{E}[d(X,Y)] \le (1-\gamma)\, d(X_0, Y_0)$ $\quad X \sim P(X_0, \cdot)$

$\qquad Y \sim P(Y_0, \cdot)$

then $T_{mix} = O\left(\frac{1}{\gamma} \log(\mathrm{diam}(T, w))\right)$

where $\mathrm{diam}(T, w) = \max_{x,y \in \Omega} d(x,y)$

(Proof) $X_0 = x \quad X_t \sim P^t(x, \cdot)$

$Y_0 \sim \pi \quad Y_t \sim \pi$

$$d_{TV}(P^t(x,\cdot), \pi) \le \Pr(X_t \ne Y_t) \le \mathbb{E}[d(X_t, Y_t)]$$

Coupling $(X_t, Y_t) \xrightarrow{P} (X_{t+1}, Y_{t+1})$

1. Take the shortest path in (T, w)

$X_t = Z_0 \cdots\cdots Z_k = Y_t \quad Z_{i-1}, Z_i \in S$

s.t. $d(X_t, Y_t) = \sum_{i=1}^{k} w(Z_{i-1}, Z_i)$

2. For $i=1$ to k:

Take a coupling of $(Z_{i-1}, Z_i) \xrightarrow{P} (W_{i-1}, W_i)$

s.t. $\mathbb{E}[d(W_{i-1}, W_i) \mid (Z_{i-1}, Z_i)] \le (1-\gamma)\, d(Z_{i-1}, Z_i)$

3. Compose these coupling to get a ground coupling,

$(Z_0, Z_1, \cdots, Z_k) \xrightarrow{P} (W_0, W_1, \cdots, W_k)$

where $W_i \sim P(Z_i, \cdot)$

Set $X_{t+1} = W_0 \quad Y_{t+1} = W_k$

$$
\begin{aligned}
\mathbb{E}[d(X_{t+1}, Y_{t+1}) \mid (X_t, Y_t)] &\le \mathbb{E}[d(w_0, w_k) \mid (z_0, z_k)] \\
&\le \sum_{i=1}^{k} \mathbb{E}[d(w_{i-1}, w_i) \mid (z_0, z_k)] \\
&\le (1-\gamma) \sum_{i=1}^{k} d(z_{i-1}, z_i) \\
&= (1-\gamma)\, d(X_t, Y_t)
\end{aligned}
$$

$$d_{TV}(P^t(x,\cdot), \pi) \le (1-\gamma)^t d(x_0, y_0) \le \frac{1}{4}$$

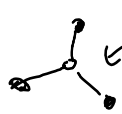
when $t \ge \frac{1}{\gamma} \log(4\, \mathrm{diam}(T, w))$

Product Space : $\Omega = [q]^n \quad n \in \mathbb{N}^+$

μ : distribution on Ω

$X_t \xrightarrow{GD} X_{t+1}$

1. Pick $i \in [n]$ u.a.r.
2. Sample $c \sim \mu_i^\sigma := \mu_i(\cdot \mid \sigma)$ where $\sigma = (X_t)_{[n] \setminus \{i\}}$



← pick coloring different from neighbors
conditioning on neighbors

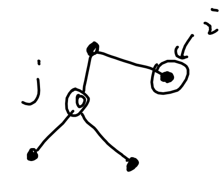
3. $X_{t+1}(j) = \begin{cases} c & \text{if } j = i \\ X_t(j) & \text{if } j \ne i \end{cases}$

Lemma : GD is reversible w.r.t. μ, so μ is stationary

(Def) Dobrushin Influence Matrix

$R \in \mathbb{R}^{n \times n}$ $\quad R(i,i) = 0 \quad \forall i \in [n]$

$$R(i,j) = \max_{\substack{(\sigma,\tau)\text{: pair of} \\ \text{config on } [n]\setminus\{j\} \\ \text{that differ only at } i}} d_{TV}(\mu_j^{\sigma}, \mu_j^{\tau}) \quad (i \neq j)$$



sample j conditioned on σ & τ respectively

(Theorem) Dob...-Shlosman uniqueness $\Rightarrow$ mixing

If $\|R\|_\infty \le 1-\delta$ $\quad$ (i.e. $\forall i \in [n]$, $\sum_{j=1}^{n} R(i,j) \le 1-\delta$)

$$T_{mix} = O\left(\frac{1}{\delta} n \log n\right)$$

You can replace $\|\cdot\|_\infty$ with $\|\cdot\|_p$

(Proof) $T = (\Omega, S)$ Hamming Graph $\quad (x,y \in S \Leftrightarrow d_H(x,y) = 1)$
unweighted

For $x, y \in \Omega$ with $d_H(x,y) = 1$, say $x_i \neq y_i$ & $x_{-i} = y_{-i}$

Construct a coupling $(x,y) \xrightarrow{GD} (X, Y)$

1. Pick $j \in [n]$ u.a.r.
2. Pick $(c_1, c_2) \in [q] \times [q]$ from an optimal coupling of $\mu_j(\cdot \mid x_{-j})$ & $\mu_j(\cdot \mid y_{-j})$
3. $X(k) = \begin{cases} c_1 & \text{if } k = j \\ x(k) & \text{if } k \neq j \end{cases}$ $\quad Y(k) = \begin{cases} c_2 & \text{if } k = j \\ y(k) & \text{if } k \neq j \end{cases}$

$$\mathbb{E}[d_H(X,Y)] - 1 \le \frac{1}{n}(-1) + \sum_{j \ne i} \frac{1}{n} R(i,j)$$

$$\le -\frac{\delta}{n}$$

Applications

Hardcore Model

$G = (V, E)$ graph of max deg Δ

$\mathcal{I}(G)$ independent sets of $G \subseteq \{0,1\}^V$

$\mu(\sigma) = \frac{1}{Z} \lambda^{|\sigma|} \quad \forall \sigma \in \mathcal{I}(G) \quad$ where $|\sigma| = \sum_{v \in V} \sigma(v)$

GD for hardcore

1. $v \in V$ u.a.r.

2. If $(\exists w \in N(v) : X_t(w) = 1)$ set $X_{t+1}(w) = 0$

   otherwise $(\forall w \in N(v),\ X_t(w) = 0)$ set $X_{t+1}(v) = \begin{cases} 1 & \text{w.p. } \frac{\lambda}{1+\lambda} \\ 0 & \text{w.p. } \frac{1}{1+\lambda} \end{cases}$

$$R(u,v) = \begin{cases} 0 & \text{if } v \notin N(u) \\ \frac{\lambda}{1+\lambda} & \text{if } v \in N(u) \end{cases}$$

So $R = \frac{\lambda}{1+\lambda} A \leftarrow$ adj matrix

$\|R\|_\infty = \frac{\lambda}{1+\lambda} \|A\|_\infty \le \frac{\lambda}{1+\lambda} \Delta \le 1 - \delta$

$\Leftrightarrow \quad \lambda \le \frac{1-\delta}{\Delta - 1 + \delta} \le \frac{1-\delta'}{\Delta - 1} \qquad T_{mix} = O(\frac{n}{\delta} \log n)$

Ising Model

$G = (V, E)$ graph of max deg Δ

$\mu(\sigma) = \frac{1}{Z} \exp(\beta \sum_{(u,v)\in E} \sigma_u \sigma_v) \quad \sigma \in \{-1, +1\}^V$

where $\beta \in \mathbb{R}$

More generally, $\mu(\sigma) = \frac{1}{Z} \exp(\frac{1}{2} \sigma^T J \sigma) \quad J \in \mathbb{R}^{n\times n} \quad \sigma \in \{-1,+1\}^n$

$R(i,j) \le \tanh|J_{ij}| \le |J_{ij}|$

$\Leftarrow \|R\|_\infty \le \|J\|_\infty$

If $J = \beta A$ then $R \le (\tanh|\beta|) A$

$\Rightarrow \|R\|_\infty \le \tanh|\beta| \cdot \Delta$

$\|J\|_\infty \le 1-\delta \Rightarrow T_{mix} = O(\frac{n}{\delta} \log n)$

Random Matching

$G = (V, E)$ a graph

$\bar{\mathcal{M}} = \bar{\mathcal{M}}(G)$ set of matchings of G

Goal: sample $M \in \bar{\mathcal{M}}$ u.a.r.

Construction of MC:

1. Choose $e = uv \in E$ u.a.r.
2. Consider four cases:

   (i) (Add) u, v are unmatched

   (ii) (Remove) u, v already matched

   (iii) (Slide) If u is unmatched & v is matched by $e' = vw \in E$

   then $X' = X_t \cup e \setminus e'$

   (iv) Otherwise $X' = X_t$

3. $X_{t+1} = \begin{cases} X' & \text{with prob } \frac{1}{2} \\ X_t & \end{cases}$

Claim: This MC is ergodic & symmetric.

stationary distribution is unif$(\bar{\mathcal{M}})$

For all $\sigma, \tau \in \Omega$, define a path γ in (Ω, P)

Observation: $\sigma \oplus \tau$ consists of alternating paths & alternating even cycles

Define the path $\sigma \rightsquigarrow \tau$

① Order components in $\sigma \oplus \tau$ by min vertex index

Given a transition:

$$\rho(M, M') = \frac{1}{\pi(M) \cdot P(M, M')} \sum_{(\sigma, \tau) \in \mathcal{P}_{MM'}} \pi(\sigma) \pi(\tau)$$

$$P(M, M') = \frac{1}{2m} \qquad = 2m \cdot \frac{|\mathcal{P}_{MM'}|}{|\bar{M}|}$$

~ try to prove $|\mathcal{P}_{MM'}| < |\bar{M}|$

construct an injection

$T_{mix} = O(m n^2 \log n)$

$\pi(M) = \frac{1}{Z} \lambda^{|M|} \quad \forall M \in \bar{M}$ (hardcore on line graph)