STATS 217: Introduction to Stochastic Processes I

Lecture 16

The Metropolis chain

- Suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.

The Metropolis chain

- Suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.
- The **Metropolis chain** provides a very general way to construct a transition matrix P with stationary distribution π.
- Moreover, the transition matrix only depends on $\tilde{\pi}$ and not π, which as we have seen, is a very important consideration.

The Metropolis chain

- Suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.
- The **Metropolis chain** provides a very general way to construct a transition matrix P with stationary distribution π.
- Moreover, the transition matrix only depends on $\tilde{\pi}$ and not π, which as we have seen, is a very important consideration.
- The Metropolis chain consists of two components:
  - The **base chain**, which is simply an irreducible Markov chain on S that we can efficiently simulate. We will denote the transition matrix of the base chain by Ψ.

The Metropolis chain

- Suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.
- The **Metropolis chain** provides a very general way to construct a transition matrix P with stationary distribution π.
- Moreover, the transition matrix only depends on $\tilde{\pi}$ and not π, which as we have seen, is a very important consideration.
- The Metropolis chain consists of two components:
  - The **base chain**, which is simply an irreducible Markov chain on S that we can efficiently simulate. We will denote the transition matrix of the base chain by Ψ. does not depend on π
  - The **Metropolis filter**, which is applied on top of the base chain to get the correct stationary distribution π.

depends only on $\tilde{\pi}$

The Metropolis chain

$\pi(x) \propto \exp(-\beta H(x))$

$H(x) = -\sum x_i x_j - h \sum x_i$

- For instance, in the case of the Ising model, for which $S = \{\pm 1\}^n$, the base chain can simply be the lazy random walk on $\{\pm 1\}^n$.

The Metropolis chain

- For instance, in the case of the Ising model, for which $S = \{\pm 1\}^n$, the base chain can simply be the lazy random walk on $\{\pm 1\}^n$.
- For now, suppose that the base chain ~~is symmetric~~ satisfies detailed balance condition with respect to the uniform distribution i.e.,

$$\Psi_{x,y} = \Psi_{y,x} \quad \forall x, y \in S.$$

In practice, this is often the case.

The Metropolis chain

- For instance, in the case of the Ising model, for which $S = \{\pm 1\}^n$, the base chain can simply be the lazy random walk on $\{\pm 1\}^n$.
- For now, suppose that the base chain is symmetric with respect to the uniform distribution i.e.,

$$\Psi_{x,y} = \Psi_{y,x} \quad \forall x, y \in S.$$

  In practice, this is often the case.
- On the homework, you will analyze a version of the Metropolis chain where the base chain is not necessarily symmetric.

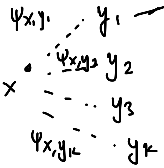
The Metropolis chain

- Given a base chain satisfying $\Psi_{x,y} = \Psi_{y,x}$, the transition matrix of the Metropolis chain is defined by

$$P_{x,y} = \begin{cases} \Psi_{x,y} \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} & y \neq x \\ 1 - \sum_{z \neq x} \Psi(x,z) \cdot \min\left\{1, \frac{\pi(z)}{\pi(x)}\right\} & y = x. \end{cases}$$

- In other words, the Metropolis filter is an acceptance-rejection filter, which accepts the proposed move of Ψ from x to y with probability $\min\{1, \pi(y)/\pi(x)\}$.

ask : is $\pi(y_1) \geq \pi(x)$?



The Metropolis chain

- Given a base chain satisfying $\Psi_{x,y} = \Psi_{y,x}$, the transition matrix of the Metropolis chain is defined by

$$P_{x,y} = \begin{cases} \Psi_{x,y} \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} & y \neq x \\ 1 - \sum_{z \neq x} \Psi(x,z) \cdot \min\left\{1, \frac{\pi(z)}{\pi(x)}\right\} & y = x. \end{cases}$$

- In other words, the Metropolis filter is an acceptance-rejection filter, which accepts the proposed move of Ψ from x to y with probability $\min\{1, \pi(y)/\pi(x)\}$.
- Thus, moves to states with greater stationary measure are always accepted, whereas moves to states with smaller stationary measure are rejected with probability $1 - (\pi(y)/\pi(x))$.

example: ising model $\pi \propto \tilde{\pi}$

$\tilde{\pi}(x) = \exp(-\beta H(x))$

base chain: lazy Random walk

$x \in \{\pm 1\}^n$

$\left\{ \begin{array}{l} \text{① choose } i \sim \{1, \dots n\} \\ \text{② choose } b \sim \{0,1\} \\ \text{③ } y = x \text{ except } i\text{th bit is } b \end{array} \right\}$

The Metropolis chain

$H(y) \le H(x)$
move to y

$H(y) > H(x)$
flip a coin heads w.p.
$0 < \frac{\exp(-\beta H(y))}{\exp(-\beta H(x))} < 1$

- Given a base chain satisfying $\Psi_{x,y} = \Psi_{y,x}$, the transition matrix of the Metropolis chain is defined by

$$P_{x,y} = \begin{cases} \Psi_{x,y} \cdot \min\left\{1, \boxed{\frac{\pi(y)}{\pi(x)}}\right\} & y \neq x \\ 1 - \sum_{z \neq x} \Psi(x,z) \cdot \min\left\{1, \boxed{\frac{\pi(z)}{\pi(x)}}\right\} & y = x. \end{cases}$$

- In other words, the Metropolis filter is an acceptance-rejection filter, which accepts the proposed move of Ψ from x to y with probability $\min\{1, \pi(y)/\pi(x)\}$.
- Thus, moves to states with greater stationary measure are always accepted, whereas moves to states with smaller stationary measure are rejected with probability $1 - (\pi(y)/\pi(x))$.
- Also, note that $\pi(x)/\pi(y) = \tilde{\pi}(x)/\tilde{\pi}(y)$.

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \leq \pi(y)$.

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \leq \pi(y)$.
- Then,

$$\pi(x) \cdot P_{x,y} = \pi(x) \cdot \Psi_{x,y} \underbrace{\min(1, \pi(y)/\pi(x))}_{= 1} = \pi(x)\Psi_{x,y}, \quad \text{and}$$

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \leq \pi(y)$.
- Then,

$$\pi(x) \cdot P_{x,y} = \pi(x) \cdot \Psi_{x,y} \min(1, \pi(y)/\pi(x)) = \underline{\underline{\pi(x)\Psi_{x,y}}}, \quad \text{and}$$
$$\underline{\underline{\pi(y) \cdot P_{y,x}}} = \cancel{\pi(y)} \cdot \underline{\underline{\Psi_{y,x}}} \underbrace{\min(1, \pi(x)/\pi(y))}_{= \frac{\pi(x)}{\cancel{\pi(y)}}}$$

$$= \pi(x)\Psi_{y,x} \underset{\text{assumed } \Psi_{x,y} = \Psi_{y,x}}{=} \underline{\underline{\pi(x)}}\,\underline{\underline{\Psi_{x,y}}}.$$

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \leq \pi(y)$.
- Then,

$$\begin{aligned}
\pi(x) \cdot P_{x,y} &= \pi(x) \cdot \Psi_{x,y} \min(1, \pi(y)/\pi(x)) = \pi(x)\Psi_{x,y}, \quad \text{and} \\
\pi(y) \cdot P_{y,x} &= \pi(y) \cdot \Psi_{y,x} \min(1, \pi(x)/\pi(y)) \\
&= \pi(y) \cdot \Psi_{x,y} \cdot \pi(x)/\pi(y)
\end{aligned}$$

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \le \pi(y)$.
- Then,

$$\begin{aligned}
\pi(x) \cdot P_{x,y} &= \pi(x) \cdot \Psi_{x,y} \min(1, \pi(y)/\pi(x)) = \pi(x)\Psi_{x,y}, \quad \text{and} \\
\pi(y) \cdot P_{y,x} &= \pi(y) \cdot \Psi_{y,x} \min(1, \pi(x)/\pi(y)) \\
&= \pi(y) \cdot \Psi_{x,y} \cdot \pi(x)/\pi(y) \\
&= \pi(x) \cdot \Psi_{x,y}
\end{aligned}$$

The Metropolis chain

- To show that π is a stationary distribution for P, it is sufficient to show that P is reversible with respect to π.
- Recall this means that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ for all $x, y \in S$.
- So, fix $x, y \in S$. Without loss of generality, suppose that $\pi(x) \leq \pi(y)$.
- Then,

$$\pi(x) \cdot P_{x,y} = \pi(x) \cdot \Psi_{x,y} \min(1, \pi(y)/\pi(x)) = \pi(x)\Psi_{x,y}, \quad \text{and}$$

$$\begin{aligned}
\pi(y) \cdot P_{y,x} &= \pi(y) \cdot \Psi_{y,x} \min(1, \pi(x)/\pi(y)) \\
&= \pi(y) \cdot \Psi_{x,y} \cdot \pi(x)/\pi(y) \\
&= \pi(x) \cdot \Psi_{x,y} \\
&= \pi(x) \cdot P_{x,y}
\end{aligned}$$

we assume

$\Psi_{x,y} = \Psi_{y,x}$

The Metropolis chain

- Therefore, we see that π is a stationary distribution of P.

The Metropolis chain

- Therefore, we see that π is a stationary distribution of P.
- Moreover, if Ψ is irreducible and $\pi(x) > 0$ for all $x \in S$, then clearly P is also irreducible.
- Therefore, in this case, π is the unique stationary distribution of P.

$$P_{x,y} = \begin{cases} \Psi_{x,y} \min\left(1, \frac{\pi(y)}{\pi(x)}\right) & y \neq x \\ \text{complement} & y = x. \end{cases}$$

for any x, z

$x \longrightarrow z$

$z \longrightarrow x$

this is true for Ψ

$x \to x_1 \to x_2 \dots \to x_\ell \to z$

s.t. $\Psi_{x,x_1} > 0 \dots \Psi_{x_\ell, z} > 0$

The Metropolis chain

- Therefore, we see that π is a stationary distribution of P.
- Moreover, if Ψ is irreducible and $\pi(x) > 0$ for all $x \in S$, then clearly P is also irreducible.
- Therefore, in this case, π is the unique stationary distribution of P.
- Also, if Ψ is aperiodic and $\pi(x) > 0$ for all $x \in S$, then clearly P is also aperiodic.

→ since Ψ is irreducible

→ know that every state has the same period

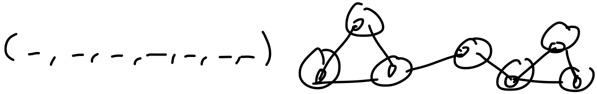
→ only need to show that some state has period 1.

The Metropolis chain

- Therefore, we see that π is a stationary distribution of P.
- Moreover, if Ψ is irreducible and $\pi(x) > 0$ for all $x \in S$, then clearly P is also irreducible.
- Therefore, in this case, π is the unique stationary distribution of P.
- Also, if Ψ is aperiodic and $\pi(x) > 0$ for all $x \in S$, then clearly P is also aperiodic.
- Therefore, by the convergence theorem, if $(X_n)_{n\geq 0}$ is a DTMC with transition matrix P and with arbitrary initial state X_0, then as $n \to \infty$, the distribution of X_n converges to π.

bad base chain: on $\{\pm 1\}^n$ stay where you are w.p. $1 - \frac{1}{2^{10^{10^{10^{10}}}}}$ and o.w. move to random neighbor.

Application to optimization

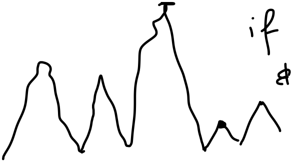


- Let $G = (V, E)$ be a graph and let $f: V \to \mathbb{R}$ be a real-valued function.
- A fundamental computational task is to find a vertex v where f is maximized.

Application to optimization

- Let $G = (V, E)$ be a graph and let $f \colon V \to \mathbb{R}$ be a real-valued function.
- A fundamental computational task is to find a vertex v where f is maximized.
- However, if V is too large, an exhaustive search may be infeasible.

- idea: greedy search.



if at x w/ value $f(x)$
& have $y \sim x$ w/ value $f(y) > f(x)$
move to y (say move to max neighbor)

"gradient ascent"

Application to optimization

- Let $G = (V, E)$ be a graph and let $f: V \to \mathbb{R}$ be a real-valued function.
- A fundamental computational task is to find a vertex v where f is maximized.
- However, if V is too large, an exhaustive search may be infeasible.
- We can use the Metropolis chain for this task. Consider the probability distribution on V given by

base chain is lazy
Random walk on graph

$$\pi_\beta(v) = \underline{\underline{e^{\beta f(v)}}}/Z(\beta),$$

x — y

where $Z(\beta) = \sum_{v \in V} e^{\beta f(v)}$ is the partition function.

y is proposed by the base chain

if $f(y) < f(x)$

accept move w.p. $\dfrac{e^{\beta f(y)}}{e^{\beta f(x)}} < 1.$

$f(y) \geq f(x)$

Application to optimization

- Let $G = (V, E)$ be a graph and let $f \colon V \to \mathbb{R}$ be a real-valued function.
- A fundamental computational task is to find a vertex v where f is maximized.
- However, if V is too large, an exhaustive search may be infeasible.
- We can use the Metropolis chain for this task. Consider the probability distribution on V given by

$$\pi_\beta(v) = e^{\beta f(v)}/Z(\beta),$$

  where $Z(\beta) = \sum_{v \in V} e^{\beta f(v)}$ is the partition function.
- Since we have access to $f(v)$, we can simulate the Metropolis chain for $\pi_\beta(v)$.

metropolis chain $\longrightarrow \pi_\beta$

Application to optimization

$\lim_{\beta \to \infty} \pi_\beta =$ uniform dist. on ~~opti~~ maximisers of f.

- The key point now is the following: let

$$V^* = \{v \in V : f(v) = f^* = \max_{u \in V} f(u)\}$$

denote the set of maximizers of f.

Application to optimization

- The key point now is the following: let

$$V^* = \{v \in V : f(v) = f^* = \max_{u \in V} f(u)\}$$

denote the set of maximizers of f.

- Then,

$$\lim_{\beta \to \infty} \pi_\beta(v) = \lim_{\beta \to \infty} \frac{e^{\beta f(v)}}{\underbrace{\underline{|V^*| e^{\beta f^*}} + \sum_{u \notin V^*} \underline{e^{\beta f(u)}}}_{Z_\beta}}$$

Application to optimization

- The key point now is the following: let

$$V^* = \{v \in V : f(v) = f^* = \max_{u \in V} f(u)\}$$

denote the set of maximizers of f.

- Then,

$$\begin{aligned} \lim_{\beta \to \infty} \pi_\beta(v) &= \lim_{\beta \to \infty} \frac{e^{\beta f(v)}}{|V^*| e^{\beta f^*} + \sum_{u \notin V^*} e^{\beta f(u)}} \\ &= \lim_{\beta \to \infty} \frac{e^{\beta(f(v) - f^*)}}{|V^*| + \underbrace{\sum_{u \notin V^*} e^{\beta(\underbrace{f(u) - f^*}_{< 0})}}_{\to 0}} \end{aligned}$$

Application to optimization

- The key point now is the following: let

$$V^* = \{v \in V : f(v) = f^* = \max_{u \in V} f(u)\}$$

denote the set of maximizers of f.

- Then,

$$\begin{aligned}
\lim_{\beta \to \infty} \pi_\beta(v) &= \lim_{\beta \to \infty} \frac{e^{\beta f(v)}}{|V^*| e^{\beta f^*} + \sum_{u \notin V^*} e^{\beta f(u)}} \\
&= \lim_{\beta \to \infty} \frac{e^{\beta (f(v) - f^*)}}{|V^*| + \sum_{u \notin V^*} e^{\beta (f(u) - f^*)}} \\
&= \frac{\mathbb{1}[v \in V^*]}{|V^*|},
\end{aligned}$$

which is the uniform distribution over the set of optimizers.

Example

We discuss a practical example of optimizing via the Metropolis chain from Persi Diaconis's article *The Markov Chain Monte Carlo Revolution*. This is drawn from course work of former Stanford students Marc Coram and Phil Beineke. All figures in these slides are from *The Markov Chain Monte Carlo Revolution*.

Example

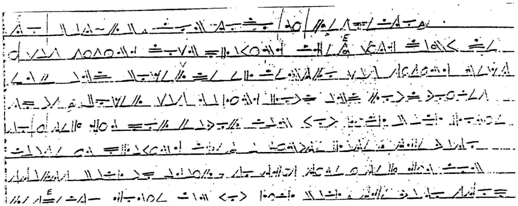


Figure 1: Coded message from state prison system, delivered by psychologist. The goal is to decode the message.

Example

A – ^ C – ʌ
B – ||

- Guess: substitution cipher i.e. the decoding function is given by

$$f : \{\text{symbols used for code}\} \to \{\text{usual alphabet}\}.$$

- Idea: define the score of the decoding function to be

$$S(f) := \prod_{i=1}^{n} \text{score}(f(\alpha_i), f(\alpha_{i+1})),$$

^ → z
|| → x
^|| ^|| ^||

where the coded message from the state prison is $\alpha_1, \dots, \alpha_n$, and for two characters x, y, score(x, y) denotes the fraction of time x and y appear successively in the English language.

- score(x, y) is determined empirically by analyzing a bunch of long texts, Wikipedia, etc.
- Attempt: Find f maximizing $S(f)$ by running the Metropolis algorithm where the base chain is given by transpositions. → in permutation, swap i and j
- Is this a reasonable strategy?

Example

ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END

Figure 2: Test run on fragment from Hamlet. This is the original version which is then encoded with a random permutation.

Example

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

$$\begin{bmatrix} 1 & 2 & 3 & 4 & \cdots & 26 & \cdots \\ x & A & Z & Y & B\ W \cdots & * & @ \ - \cdots \end{bmatrix}$$

Figure 3: Performance of the Metropolis algorithm for Hamlet.

Example

Decoding returned by the Metropolis algorithm on the prison text after a few thousand steps (+ some manual cleanup)

```
THE F**K UP CAUSE ZM TILEH OF YOUL VOICE AND IF YOU GOT A PLOBREM WITG
IT WE CAN GO TY CERDA AND IFNDRE IT I LEALY FERT DITLESPECTED THATS
WHY I TORD HIM ANYWAS AFTEL I TER HIM THAT THE NEXT THING I KNOW THE
VATO SRASHES ME AND REAVES DY THE TIME I FIGULE IM HIT I TLY TO ?ET
AWAY DUT THE C O IS WARKING IN MY DILECTION AND NA ?ETS ME LIGHT DY A
CERDA SO I GO TO THE HORE WHEN ?M IN THE HORE ?Y HOME BOYS HIT DOXEL
SO NOW B I? AHSO IN THE HORE WHIRE IM IN THE HORE IM GESONG SCHORD
WLONG AND GE?ONG TORD THAT YOU DONT DLING PELSONAR PEDO INTO THE HORE
ANYWAYS ?? AND ? END ?P O? THE SAME YALDA AND ON OUL YALDA THE LE I?
ONDY ? OF US ?ND ?? NEGLO? OUT OF THEM ?? NEGLOS ? OL ? OF THEM ALE IM
THE HORE FOL ? LIOT WITH THE LAJA AND OUT OF US ? ? OF US ALE IN THE
HORE FOL THE SAME LIOT ANYWAWAYS I TAKE ONE ON ? AND REAVE IT AT THAT
AND TER HIM WELE GONA CER IT UP I WANTED TO CER IT UP WITH ? BECAUSE
HE WAS ?OING???HOM? XITHIN ???? DAYS AND I WAN TED TO GET MINE??BEFOLE
HE HEFT IM NOT SULE BUT ?E CERD UP THE NEXT DAY?AND
```