Computationally Efficient Estimation of the Spectral Gap of a Markov Chain

Richard Combes
joint work with Mikael Touati

UCL 2019

CentraleSupélec

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

The problem

- Data Markov Chain $(X_t)_t$ on a finite state space Ω
- Transition matrix P reversible and lazy with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|} \geq 0$
- Simulation methods P unknown, and we may only simulate sample paths.
- Efficient algorithms $|\Omega|$ is large, only $o(|\Omega|)$ memory available
- We are interested in confidence intervals, not point estimates

Goal: Estimate the spectral gap of an unknown Markov Chain in a computationally efficient manner.

Spectral gap and mixing time

- Purpose The spectral gap quantifies mixing time

$$\frac{1}{2}\max_{\mu}\sum_{x\in\Omega}\left|\mathbb{P}(X_{t_m}=x|X_0\sim\mu)-\pi(x)\right|\leq\frac{1}{4}$$

- Mixing time is upper bounded by relaxation time $t_r=\frac{1}{1-\lambda_2}$

$$(t_r-1)\ln 2\leq t_m\leq t_r\ln\frac{4}{\min_{x\in\Omega}\pi(x)}.$$

- Bounds upper bound the relaxation time to ensure mixing
- Applications MCMC, Bayesian Statistics, Statistical Physics, Queuing Systems etc.

Spectral gap and mixing time

- Purpose The spectral gap quantifies mixing time

$$\frac{1}{2}\max_{\mu}\sum_{x\in\Omega}\left|\mathbb{P}(X_{t_m}=x|X_0\sim\mu)-\pi(x)\right|\leq\frac{1}{4}$$

- Mixing time is upper bounded by relaxation time $t_r = \frac{1}{1-\lambda_2}$

$$(t_r-1)\ln 2 \leq t_m \leq t_r \ln\frac{4}{\min_{x\in\Omega}\pi(x)}.$$

- Bounds upper bound the relaxation time to ensure mixing
- Applications MCMC, Bayesian Statistics, Statistical Physics, Queuing Systems etc.

Spectral gap and mixing time

- Purpose The spectral gap quantifies mixing time

$$\frac{1}{2}\max_{\mu}\sum_{x\in\Omega}\left|\mathbb{P}(X_{t_m}=x|X_0\sim\mu)-\pi(x)\right|\leq\frac{1}{4}$$

- Mixing time is upper bounded by relaxation time $t_r=\frac{1}{1-\lambda_2}$

$$(t_r-1)\ln 2\leq t_m\leq t_r\ln\frac{4}{\min_{x\in\Omega}\pi(x)}.$$

- Bounds upper bound the relaxation time to ensure mixing
- Applications MCMC, Bayesian Statistics, Statistical Physics, Queuing Systems etc.

Spectral gap and mixing time

- Purpose The spectral gap quantifies mixing time

$$\frac{1}{2}\max_{\mu}\sum_{x\in\Omega}\left|\mathbb{P}(X_{t_m}=x|X_0\sim\mu)-\pi(x)\right|\leq\frac{1}{4}$$

- Mixing time is upper bounded by relaxation time $t_r = \frac{1}{1-\lambda_2}$

$$(t_r-1)\ln 2\leq t_m\leq t_r\ln\frac{4}{\min_{x\in\Omega}\pi(x)}.$$

- Bounds upper bound the relaxation time to ensure mixing
- Applications MCMC, Bayesian Statistics, Statistical Physics, Queuing Systems etc.

Computational models

- Random Transition Function A black box `NextState`(x) simulates transitions
- Unique Sample Path A unique sample path $X_1, \ldots, X_n$.
- Reduction In fact we show that USP can be reduced to RTF without loss.

Computational models

- Random Transition Function A black box `NextState`(x) simulates transitions
- Unique Sample Path A unique sample path $X_1, \ldots, X_n$.
- Reduction In fact we show that USP can be reduced to RTF without loss.

Computational models

- Random Transition Function A black box `NextState`(x) simulates transitions
- Unique Sample Path A unique sample path $X_1, \ldots, X_n$.
- Reduction In fact we show that USP can be reduced to RTF without loss.

Trace of Powers of P

Fundamental relation:

$$\lim_{k\to\infty}\left[\sum_{x\in\Omega} P^k(x,x) - 1\right]^{\frac{1}{k}} = \lim_{k\to\infty}\left[\sum_{i=2}^{|\Omega|} \lambda_i^k\right]^{\frac{1}{k}} = \lambda_2.$$

Proposition

If U is uniformly distributed on Ω, then:

$$\lambda_2 = \lim_{k\to\infty}\left[|\Omega|\mathbb{P}(X_k = X_0 | X_0 = U) - 1\right]^{\frac{1}{k}}.$$

We use the notation:

$$m_k = \mathbb{P}(X_k = X_0|X_0 = U) = \frac{1}{|\Omega|}\sum_{x\in\Omega} P^k(x,x) = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\lambda_i^k.$$

Trace of Powers of P

Fundamental relation:

$$\lim_{k\to\infty}\left[\sum_{x\in\Omega}P^k(x,x)-1\right]^{\frac{1}{k}}=\lim_{k\to\infty}\left[\sum_{i=2}^{|\Omega|}\lambda_i^k\right]^{\frac{1}{k}}=\lambda_2.$$

Proposition

If U is uniformly distributed on Ω, then:

$$\lambda_2=\lim_{k\to\infty}\Big[|\Omega|\mathbb{P}(X_k=X_0|X_0=U)-1\Big]^{\frac{1}{k}}.$$

We use the notation:

$$m_k=\mathbb{P}(X_k=X_0|X_0=U)=\frac{1}{|\Omega|}\sum_{x\in\Omega}P^k(x,x)=\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\lambda_i^k.$$

Trace of Powers of P

Fundamental relation:

$$\lim_{k\to\infty}\left[\sum_{x\in\Omega}P^k(x,x)-1\right]^{\frac{1}{k}}=\lim_{k\to\infty}\left[\sum_{i=2}^{|\Omega|}\lambda_i^k\right]^{\frac{1}{k}}=\lambda_2.$$

Proposition

If U is uniformly distributed on Ω, then:

$$\lambda_2=\lim_{k\to\infty}\Big[|\Omega|\mathbb{P}(X_k=X_0|X_0=U)-1\Big]^{\frac{1}{k}}.$$

We use the notation:

$$m_k=\mathbb{P}(X_k=X_0|X_0=U)=\frac{1}{|\Omega|}\sum_{x\in\Omega}P^k(x,x)=\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\lambda_i^k.$$

Trace-based Estimation

An unbiased estimate for m_k is:

$$\hat{m}_k = \frac{1}{I}\sum_{j=1}^{I} \mathbf{1}\{X_k^j = X_0^j\}.$$

and we estimate λ_2 with the plug-in estimator:

$$\hat{q}_k = \Big[|\Omega|\hat{m}_k - 1\Big]^{\frac{1}{k}}.$$

Asymptotic behavior of the estimate:

$$\sqrt{I}\Big(\hat{q}_k - \Big[|\Omega|m_k - 1\Big]^{\frac{1}{k}}\Big) \underset{I\to\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}\Big(0, \frac{m_k(1-m_k)|\Omega|^2}{k^2}\Big[|\Omega|m_k - 1\Big]^{\frac{2(1-k)}{k}}\Big)$$

Problem: How to choose k to minimize error?

Trace-based Estimation

An unbiased estimate for m_k is:

$$\hat{m}_k = \frac{1}{I}\sum_{j=1}^{I} \mathbf{1}\{X_k^j = X_0^j\}.$$

and we estimate λ_2 with the plug-in estimator:

$$\hat{q}_k = \left[|\Omega|\hat{m}_k - 1\right]^{\frac{1}{k}}.$$

Asymptotic behavior of the estimate:

$$\sqrt{I}\left(\hat{q}_k - \left[|\Omega|m_k - 1\right]^{\frac{1}{k}}\right) \xrightarrow[I\to\infty]{(d)} \mathcal{N}\left(0, \frac{m_k(1-m_k)|\Omega|^2}{k^2}\left[|\Omega|m_k - 1\right]^{\frac{2(1-k)}{k}}\right)$$

Problem: How to choose k to minimize error?

Trace-based Estimation

An unbiased estimate for m_k is:

$$\hat{m}_k = \frac{1}{I} \sum_{j=1}^{I} \mathbf{1}\{X_k^j = X_0^j\}.$$

and we estimate λ_2 with the plug-in estimator:

$$\hat{q}_k = \left[|\Omega| \hat{m}_k - 1 \right]^{\frac{1}{k}}.$$

Asymptotic behavior of the estimate:

$$\sqrt{I}\left(\hat{q}_k - \left[|\Omega| m_k - 1 \right]^{\frac{1}{k}} \right) \underset{I \to \infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}\left(0, \frac{m_k(1 - m_k)|\Omega|^2}{k^2} \left[|\Omega| m_k - 1 \right]^{\frac{2(1-k)}{k}} \right)$$

Problem: How to choose k to minimize error?

Trace-based Estimation

An unbiased estimate for m_k is:

$$\hat{m}_k = \frac{1}{I} \sum_{j=1}^{I} \mathbf{1}\{X_k^j = X_0^j\}.$$

and we estimate λ_2 with the plug-in estimator:

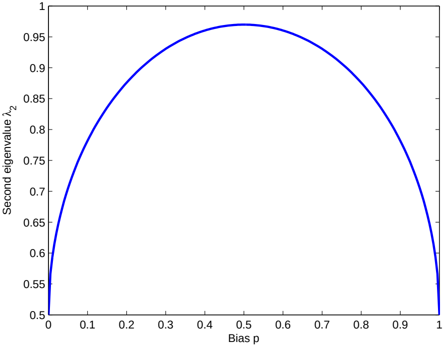
$$\hat{q}_k = \left[|\Omega| \hat{m}_k - 1 \right]^{\frac{1}{k}}.$$

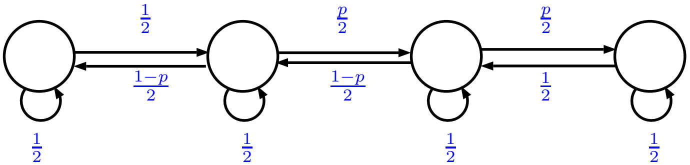
Asymptotic behavior of the estimate:

$$\sqrt{I}\left(\hat{q}_k - \left[|\Omega| m_k - 1 \right]^{\frac{1}{k}} \right) \underset{I \to \infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}\left(0, \frac{m_k(1 - m_k)|\Omega|^2}{k^2} \left[|\Omega| m_k - 1 \right]^{\frac{2(1-k)}{k}} \right)$$

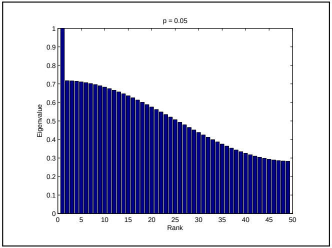
Problem: How to choose k to minimize error?

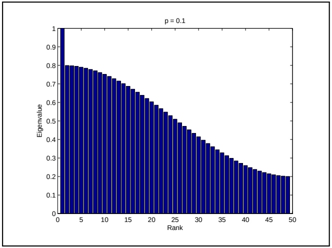
An Example: Biased Random Walk On a Line

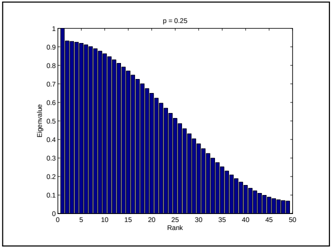


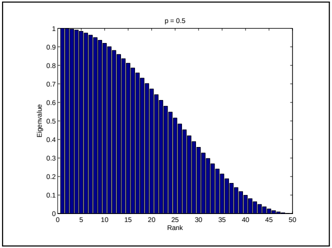


Biased Random Walk On a Line: Spectrum

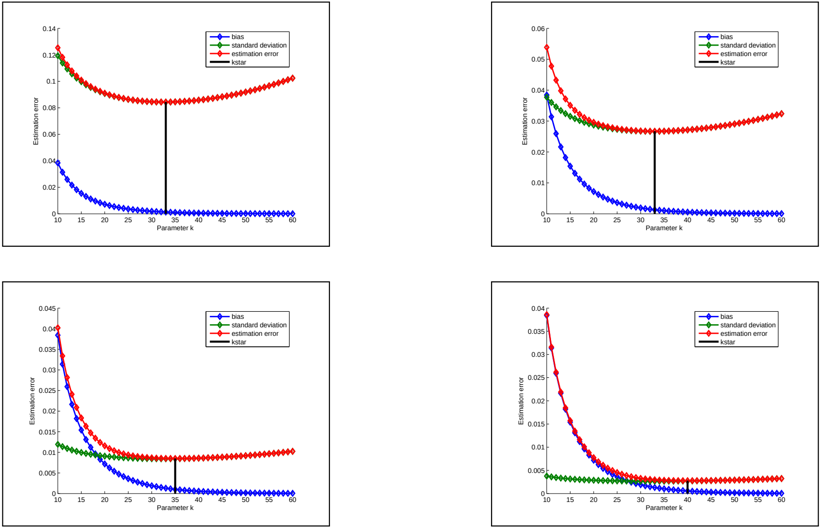




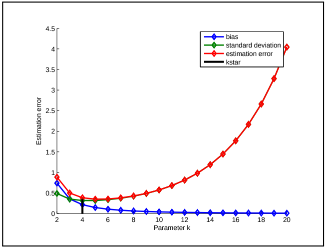


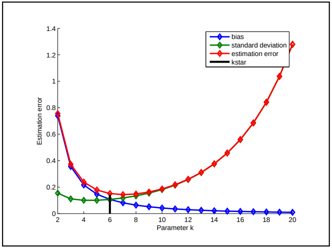


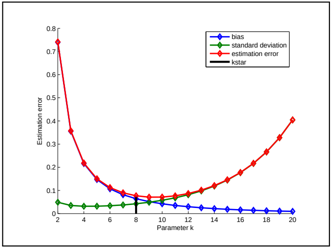
Bias Variance Tradeoff, $p = \frac{1}{2}$

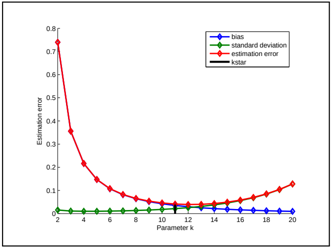


Bias Variance Tradeoff, $p = \frac{1}{10}$









UCPI: Upper Confidence Power Iteration

Algorithm 1 `UCPI` (Upper Confidence Power Iteration)

Input: P transition matrix, I number of paths, K max. length of paths , δ confidence parameter

Initialize $\hat{m}_k \leftarrow 0$, for $k = 1, \ldots, K$;

for $i = 1$ **to** I **do**

  $X_0 \leftarrow \texttt{Uniform}(\Omega)$; $X \leftarrow X_0$;

  for $k = 1$ **to** K **do**

    $X \leftarrow \texttt{NextState}(X)$; $\hat{m}_k \leftarrow \frac{i-1}{i}\hat{m}_k + \frac{1}{i}\mathbf{1}\{X_0 = X\}$;

  end for

end for

for $k = 1$ **to** K **do**

  $\hat{u}_k \leftarrow \texttt{CB}(\hat{m}_k, \frac{\delta}{2K})$; $\hat{\ell}_k \leftarrow \min\left((|\Omega|\hat{u}_k - 1)^{\frac{1}{k}}, 1\right)$;

end for

$\hat{\ell}_\star \leftarrow \min_{k=1,\ldots,K} \hat{\ell}_k$;

Output: $\hat{\ell}_\star$ an estimate and confidence upper bound for $\lambda_\star$.

Analysis of UCPI

Define the parameter

$$r = \frac{\ln 1/\lambda_2}{\ln 1/\lambda_3} \in [0, 1],$$

Theorem

Consider $\hat{\ell}_\star$ the output of `UCPI` *with parameters $\delta = \frac{1}{\sqrt{n}}$ and $K = (\ln n)^2$ and $I = \frac{n}{K} = \frac{n}{(\ln n)^2}$.*
Then:

$$\mathbb{E}|\hat{\ell}_\star - \lambda_2| \leq C \frac{|\Omega|^{\frac{r+1}{2}} \ln(1/\lambda_3)}{(\ln n)^{\frac{3r-1}{2}} n^{\frac{1-r}{2}}} \quad , \quad \forall n \geq n_0(|\Omega|, \lambda_2, \lambda_3),$$

with $C > 0$ a universal constant.

Analysis of UCPI

Define the parameter

$$r = \frac{\ln 1/\lambda_2}{\ln 1/\lambda_3} \in [0, 1],$$

Theorem

Consider $\hat{\ell}_\star$ the output of `UCPI` *with parameters $\delta = \frac{1}{\sqrt{n}}$ and $K = (\ln n)^2$ and $I = \frac{n}{K} = \frac{n}{(\ln n)^2}$.*
Then:

$$\mathbb{E}|\hat{\ell}_\star - \lambda_2| \leq C \frac{|\Omega|^{\frac{r+1}{2}} \ln(1/\lambda_3)}{(\ln n)^{\frac{3r-1}{2}} n^{\frac{1-r}{2}}} \ , \ \forall n \geq n_0(|\Omega|, \lambda_2, \lambda_3),$$

with $C > 0$ a universal constant.

Numerical Experiments (1)

$\lvert\Omega\rvert = 20, p = 0.5, \lambda^\star = 0.99, t_r = 100$				
n	UCPI $\hat{\ell}_\star$	UCPI $\hat{t}_r$	UCI $\hat{\ell}_\star$	UCI $\hat{t}_r$
10^4	1	∞	1	∞
10^5	1	∞	1	∞
10^6	0.996	272	1	∞
10^7	0.994	176	1	∞
10^8	0.993	154	1	∞
$\lvert\Omega\rvert = 20, p = 0.7, \lambda^\star = 0.95, t_r = 20$				
n	UCPI $\hat{\ell}_\star$	UCPI $\hat{t}_r$	UCI $\hat{\ell}_\star$	UCI $\hat{t}_r$
10^4	1	∞	1	∞
10^5	0.993	151	1	∞
10^6	0.977	45	1	∞
10^7	0.969	32	1	∞
10^8	0.963	27	1	∞
$\lvert\Omega\rvert = 20, p = 0.9, \lambda^\star = 0.8, t_r = 5$				
n	UCPI $\hat{\ell}_\star$	UCPI $\hat{t}_r$	UCI $\hat{\ell}_\star$	UCI $\hat{t}_r$
10^4	1	∞	1	∞
10^5	0.985	69	1	∞
10^6	0.939	16	1	∞
10^7	0.899	10	1	∞
10^8	0.875	8	1	∞

Table: Biased random walk on a line.

Numerical Experiments (2)

$\lvert\Omega\rvert = 100$, $d = 5$, $\lambda^\star = 0.88$, $t_r = 8.3$				
n	UCPI $\hat{\ell}_\star$	UCPI $\hat{t}_r$	UCI $\hat{\ell}_\star$	UCI $\hat{t}_r$
10^4	1	∞	1	∞
10^5	1	∞	1	∞
10^6	0.979	49	1	∞
10^7	0.958	24	1	∞
10^8	0.934	15	1	∞
$\lvert\Omega\rvert = 100$, $d = 10$, $\lambda^\star = 0.78$, $t_r = 4.5$				
n	UCPI $\hat{\ell}_\star$	UCPI $\hat{t}_r$	UCI $\hat{\ell}_\star$	UCI $\hat{t}_r$
10^4	1	∞	1	∞
10^5	0.962	26	1	∞
10^6	0.980	47	1	∞
10^7	0.941	17	1	∞
10^8	0.902	10	1	∞

Table: Random walk on a d-regular graph.

Numerical Experiments (3)

$n = 10^6$	UCPI $\mathbb{P}(\hat{\ell}_\star < 1)$	UCI $\mathbb{P}(\hat{\ell}_\star < 1)$
Case 1 ($p = 0.5$)	1	0
Case 1 ($p = 0.7$)	1	0
Case 1 ($p = 0.9$)	1	0
Case 2 ($d = 5$)	1	0
Case 2 ($d = 10$)	1	0

Table: Informativeness of confidence upper bounds.

Extensions

What about `UCPI` in more general settings ?

- RTF model with n samples reduces to USP model with

$$n\frac{\ln n}{\ln n + \max_{x,y}\xi(x,y)} \underset{n\to\infty}{\sim} n$$

samples, $\xi(x,y)$ - expected hitting time from x to y.

- Non-uniform starting distributions can be handled using

$$\lambda_2 = \lim_{k\to\infty}\left[\mathbb{E}_{U\sim\mu}\left(\frac{\mathbf{1}\{X_k = X_0\}}{\mu(U)}|X_0 = U\right) - 1\right]^{\frac{1}{k}}.$$

- Non-lazy chains can be handled by applying `UCPI` to P^2.

Extensions

What about UCPI in more general settings ?

- RTF model with n samples reduces to USP model with

$$n\frac{\ln n}{\ln n + \max_{x,y}\xi(x,y)} \underset{n\to\infty}{\sim} n$$

samples, $\xi(x,y)$ - expected hitting time from x to y.

- Non-uniform starting distributions can be handled using

$$\lambda_2 = \lim_{k\to\infty}\left[\mathbb{E}_{U\sim\mu}\left(\frac{\mathbf{1}\{X_k = X_0\}}{\mu(U)}|X_0 = U\right) - 1\right]^{\frac{1}{k}}.$$

- Non-lazy chains can be handled by applying UCPI to P^2.

Extensions

What about `UCPI` in more general settings ?

- RTF model with n samples reduces to USP model with

$$n\frac{\ln n}{\ln n + \max_{x,y}\xi(x,y)} \underset{n\to\infty}{\sim} n$$

samples, $\xi(x,y)$ - expected hitting time from x to y.

- Non-uniform starting distributions can be handled using

$$\lambda_2 = \lim_{k\to\infty}\left[\mathbb{E}_{U\sim\mu}\left(\frac{\mathbf{1}\{X_k = X_0\}}{\mu(U)}|X_0 = U\right) - 1\right]^{\frac{1}{k}}.$$

- Non-lazy chains can be handled by applying `UCPI` to P^2.

Extensions

What about `UCPI` in more general settings ?

- RTF model with n samples reduces to USP model with

$$n\frac{\ln n}{\ln n + \max_{x,y}\xi(x,y)} \underset{n\to\infty}{\sim} n$$

  samples, $\xi(x,y)$ - expected hitting time from x to y.

- Non-uniform starting distributions can be handled using

$$\lambda_2 = \lim_{k\to\infty}\left[\mathbb{E}_{U\sim\mu}\left(\frac{\mathbf{1}\{X_k = X_0\}}{\mu(U)}|X_0 = U\right) - 1\right]^{\frac{1}{k}}.$$

- Non-lazy chains can be handled by applying `UCPI` to P^2.

Thank you for your attention !