A Deterministic Algorithm for the Capacity of Finite-State Channels

Chengyu Wu[1], Guangyue Han[1] and Brian Marcus[2]

[1]The University of Hong Kong
[2]University of British Columbia

August, 2019

Summary

- Channel Model
- Optimization Problem with Line Search Method
- Our Algorithm
- Convergence Analysis
- Applications
- Generalization to Non-Concave Case

Channel Model

We focus on finite-state channels with input constraints. To formulate this channel, we first introduce the following notation.

- For any $F \subseteq \mathcal{X}^2$ (forbidden set) and $\delta > 0$, define

$$\Pi_{F,\delta} = \{\text{stochastic matrix } A : A_{ij} = 0, \quad \text{for } (i,j) \in F \\ \text{and } A_{ij} \geq \delta \quad \text{otherwise}\}.$$

Channel Model

We focus on finite-state channels with input constraints. To formulate this channel, we first introduce the following notation.

- For any $F \subseteq \mathcal{X}^2$ (forbidden set) and $\delta > 0$, define

$$\Pi_{F,\delta} = \{\text{stochastic matrix } A : A_{ij} = 0, \quad \text{for } (i,j) \in F \\ \text{and } A_{ij} \geq \delta \quad \text{otherwise}\}.$$

- One typical example is given by $\mathcal{X} = \{0,1\}, F = \{11\}$, i.e., the block 11 is forbidden for all binary sequences. ($(1,\infty)$-RLL constraint)

Channel Model

We are concerned with finite-state channels such that:

(a) X is an irreduaible Markov chain and there exist $F \subseteq \mathcal{X}^2$ and $\delta > 0$ such that the transition probability matrix of X belongs to $\Pi_{F,\delta}$.

Channel Model

We are concerned with finite-state channels such that:

(a) X is an irreduaible Markov chain and there exist $F \subseteq \mathcal{X}^2$ and $\delta > 0$ such that the transition probability matrix of X belongs to $\Pi_{F,\delta}$.

(b) (X, S) is a stationary Markov chain and

$$p(x_n, s_n | x_{n-1}, s_{n-1}) = p(x_n | x_{n-1}) p(s_n | x_n, s_{n-1})$$

for $n = 1, 2, \ldots$ where $p(s_n | x_n, s_{n-1}) > 0$ for any s_{n-1}, s_n, x_n.

Channel Model

We are concerned with finite-state channels such that:

(a) X is an irreduaible Markov chain and there exist $F \subseteq \mathcal{X}^2$ and $\delta > 0$ such that the transition probability matrix of X belongs to $\Pi_{F,\delta}$.

(b) (X, S) is a stationary Markov chain and

$$p(x_n, s_n | x_{n-1}, s_{n-1}) = p(x_n | x_{n-1}) p(s_n | x_n, s_{n-1})$$

for $n = 1, 2, \ldots$ where $p(s_n | x_n, s_{n-1}) > 0$ for any s_{n-1}, s_n, x_n.

(c) the channel is stationary and characterized by

$$p(y_n | y_1^{n-1}, x_1^n, s_1^{n-1}) = p(y_n | x_n, s_{n-1}) > 0$$

for $n = 1, 2, \ldots$.

Properties of the Channel

For finite-state channels satisfying (a), (b) and (c), the following can be readily verified:

1. The channel is indecomposable.

Properties of the Channel

For finite-state channels satisfying (a), (b) and (c), the following can be readily verified:

1. The channel is indecomposable.
2. Finding the capacity corresponds to solving the following optimization problem:

$$\begin{aligned} C &= \sup I(X;Y) \\ &= \sup \lim_{n\to\infty} \frac{H(Y_1^n) + H(X_1^n) - H(X_1^n, Y_1^n)}{n} \\ &= \sup \lim_{n\to\infty} H(X_1^2) + H(Y_n|Y_1^{n-1}) - H(X_n, Y_n|X_1^{n-1}, Y_1^{n-1}) \end{aligned}$$

where sup is taken over all distributions of the input $\{X_n\}_{n=1}^{\infty}$.

Properties of the Channel (Cont)

3. It has been proved (Han, 2015) that $H(Y_n|Y_1^{n-1})$ and $H(X_n, Y_n|X_1^{n-1}, Y_1^{n-1})$ converges exponentially. Hence, if we assume that the input Markov chain is parameterized by $\theta \in \Theta$ and let

$$f(\theta) := \sup_{\theta} I(X;Y)$$

$$f_k(\theta) = H(X_1^2) + H(Y_n|Y_1^{n-1}) - H(X_n, Y_n|X_1^{n-1}, Y_1^{n-1})$$

then there exist $N > 0$ and $0 < \rho < 1$ such that for $l = 0, 1, 2$,

$$||f_k^{(\ell)}(\theta) - f_{k-1}^{(\ell)}(\theta)||_2 \leq N\rho^k, \quad ||f_k^{(\ell)}(\theta) - f^{(\ell)}(\theta)||_2 \leq N\rho^k \quad (1)$$

where $f^{(l)}$ is the l-th order derivative.

Properties of the Channel (Cont)

3. It has been proved (Han, 2015) that $H(Y_n|Y_1^{n-1})$ and $H(X_n, Y_n|X_1^{n-1}, Y_1^{n-1})$ converges exponentially. Hence, if we assume that the input Markov chain is parameterized by $\theta \in \Theta$ and let

$$f(\theta) := \sup_{\theta} I(X;Y)$$

$$f_k(\theta) = H(X_1^2) + H(Y_n|Y_1^{n-1}) - H(X_n, Y_n|X_1^{n-1}, Y_1^{n-1})$$

then there exist $N > 0$ and $0 < \rho < 1$ such that for $l = 0, 1, 2$,

$$||f_k^{(\ell)}(\theta) - f_{k-1}^{(\ell)}(\theta)||_2 \leq N\rho^k, \quad ||f_k^{(\ell)}(\theta) - f^{(\ell)}(\theta)||_2 \leq N\rho^k \quad (1)$$

where $f^{(l)}$ is the l-th order derivative. The capacity can be approximated exponentially.

Classical Descent Method with Backtracking Line Search

For finding the maximum of $f(x)$, one popular method is the well-know descent method:

Descent Method

Choose a starting point $x_0 \in S$. Repeat

1. Choose a direction Δx such that

$$\Delta x \cdot \nabla f(x) > 0;$$

2. choose a step size $t > 0$;
3. update the point $x := x + t\Delta x$

until the stopping criterion is satisfied.

Classical Descent Method with Backtracking Line Search

For finding the maximum of $f(x)$, one popular method is the well-know descent method:

Descent Method

Choose a starting point $x_0 \in S$. Repeat

1. Choose a direction Δx such that

$$\Delta x \cdot \nabla f(x) > 0;$$

2. choose a step size $t > 0$;
3. update the point $x := x + t\Delta x$

until the stopping criterion is satisfied.

• When $\Delta x = \nabla f(x)$, then it is the well-known gradient descent method.

Classical Descent Method with Backtracking Line Search

One of the method to choose the step size is given by the **backtracking line search**:

Classical Descent Method with Backtracking Line Search

One of the method to choose the step size is given by the **backtracking line search**:

Backtracking line search

Let $t = 1$. For a fixed descent direction Δx, choose $0 < \alpha < 0.5, 0 < \beta < 1$ and perform

$$t := \beta t$$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$.

Backtracking Line Search

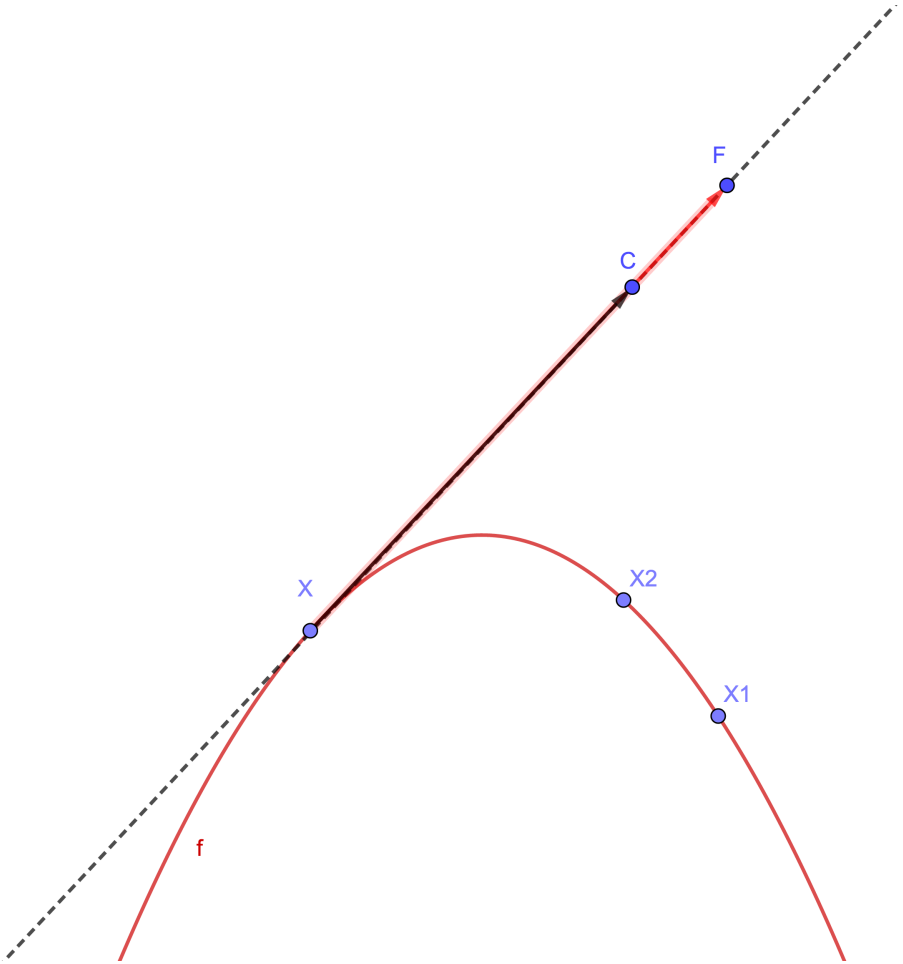

Figure: backtracking line search

Limitations of Classical Descent Methods

The classical descent methods will have trouble when:

- No explicit formula for the target function;
- the domain of the variable is not $\mathbb{R}^n$ (the convergence analysis may be very complicated).

Limitations of Classical Descent Methods

The classical descent methods will have trouble when:

- No explicit formula for the target function;
- the domain of the variable is not $\mathbb{R}^n$ (the convergence analysis may be very complicated).

In our case, however,

- $I(X(\theta); Y(\theta)) = f(\theta) = \lim_{n\to} f_k(\theta)$;
- $\theta \in \Theta$ usually a strict subset of $\mathbb{R}^n$.

Limitations of Classical Descent Methods

The classical descent methods will have trouble when:

- No explicit formula for the target function;
- the domain of the variable is not $\mathbb{R}^n$ (the convergence analysis may be very complicated).

In our case, however,

- $I(X(\theta); Y(\theta)) = f(\theta) = \lim_{n\to} f_k(\theta)$;
- $\theta \in \Theta$ usually a strict subset of $\mathbb{R}^n$.

So the classical descent method fails.
$\rightarrow$ Exponential convergence may allow us to modify it.

Our Algorithm

Algorithm 1

Step 0. Set $k = 0$, and choose $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$ and $\theta_0 \in \Theta$ such that $\nabla f_0(\theta_0) \neq 0$.

Step 1. Set $t = 1$ and increase k by 1.

Step 2. If $\nabla f_{k-1}(\theta_{k-1}) = 0$, set

$$\tau = \theta_{k-1} + t\nabla f_{k-1}(\theta_{k-1} + \rho^{k-1}),$$

otherwise, set

$$\tau = \theta_{k-1} + t\nabla f_{k-1}(\theta_{k-1}).$$

If $\tau \notin \Theta$ or

$$f_k(\tau) < f_k(\theta_{k-1}) + \alpha t||\nabla f_{k-1}(\theta_{k-1})||_2^2 - (N+M)Mt\rho^{k-1},$$

set $t = \beta t$ and go to Step 2, otherwise set $\theta_k = \tau$ and go to Step 1. (Remark: M is the upperbound on the derivatives of f.)

• Difficulty for the convergence analysis: for any k, in order to obtain a new iterate θ_{k+1} from θ_k, how many time of Step 2 is executed?

In order to solve this problem, let

- k be the number that Step 1 has been executed;
- n be the number that Step 2 has been executed.

We can rewrite our algorithm as follows:

Algorithm 1': (An equivalent form of Algorithm 1.)

Step 0. Set $n = 0, k = 0, \hat{f}_0 = f_0$, choose $\alpha \in (0, 0.5), \beta \in (0, 1)$ and $\hat{\theta}_0 \in \Theta$ such that $\nabla \hat{f}_0(\hat{\theta}_0) \neq 0$.

Step 1. Set $t = 1$ and increase k by 1.

Step 2. Increase n by 1. If $\nabla \hat{f}_{n-1}(\hat{\theta}_{n-1}) = 0$, set

$$\tau = \hat{\theta}_{n-1} + t\nabla \hat{f}_{n-1}(\hat{\theta}_{n-1} + \rho^{k-1}),$$

otherwise, set

$$\tau = \hat{\theta}_{n-1} + t\nabla \hat{f}_{n-1}(\hat{\theta}_{n-1}).$$

If $\tau \notin \Theta$ or

$$f_k(\tau) < f_k(\hat{\theta}_{n-1}) + \alpha t||\nabla \hat{f}_{n-1}(\hat{\theta}_{n-1})||_2^2 - (N+M)Mt\rho^{k-1},$$

then set $\hat{\theta}_n = \hat{\theta}_{n-1}, \hat{f}_n = f_{k-1}, t = \beta t$ and go to Step 2, otherwise, set $\hat{\theta}_n = \tau, \hat{f}_n = f_k$ and go to Step 1.

Assumptions on the Initial Point

Before stating the convergence result of Algorithm 1', we need the following observation:

For $f_k^{(l)}(\theta) \to f^{(l)}(\theta)$ exponentially and strongly concave f, suppose f has a unique maximum point that is away from the boundary of the open connected domain Θ, then we can always choose k_0 and y_0 such that, by defining

$$B := \{x : f_{k_0}(x) \geq y_0\},$$

we have B is convex and $B \subseteq \Theta$.

Convergence Result

Theorem 1

Let $f(\theta)$ and $\{f_k(\theta)\}$ have the exponential convergence properties in (1). Suppose $f(\theta)$ is strongly concave, that is, there exists $m > 0$ such that for all $\theta \in \Theta$ (open, connect),

$$\nabla^2 f(\theta) \preceq -m I_d,$$

where I_d denotes the $d \times d$-dimensional identity matrix, and moreover, $f(\theta)$ achieves its maximum at θ^* which has a positive distance to $\partial\Theta$. Then, by choosing θ_{k_0} in B and running Algorithm 1', there exist $\hat{M} > 0$ and $0 < \hat{\xi} < 1$ such that for all n,

$$|\hat{f}_n(\hat{\theta}_n) - f(\theta^*)| \leq \hat{M}\hat{\xi}^n.$$

Outline of the Proof

Suppose we are at θ_{k-1} now. Remember that

$$\tau = \theta_{k-1} + t\nabla f_{k-1}(\theta_{k-1}).$$

Define:

- $T_1(k)$: time used to satisfy $\tau \in \Theta$;
- $T_2(k)$: time used to satisfy the "increasing condition"

$$f_k(\tau) < f_k(\theta_{k-1}) + \alpha t||\nabla f_{k-1}(\theta_{k-1})||_2^2 - (N+M)Mt\rho^{k-1}$$

Outline of the Proof

Suppose we are at θ_{k-1} now. Remember that

$$\tau = \theta_{k-1} + t\nabla f_{k-1}(\theta_{k-1}).$$

Define:

- $T_1(k)$: time used to satisfy $\tau \in \Theta$;
- $T_2(k)$: time used to satisfy the "increasing condition"

$$f_k(\tau) < f_k(\theta_{k-1}) + \alpha t||\nabla f_{k-1}(\theta_{k-1})||_2^2 - (N+M)Mt\rho^{k-1}$$

Want: Uniform boundedness of $T_1(k)$, $T_2(k)$ over k.

Outline of the Proof (Cont)

Most important fact: we can treat $T_1(k)$ and $T_2(k)$ separately, i.e., first consider whether $\tau \in \Theta$, if not, iterate until $\tau \in \Theta$; after this is satisfied, consider the "increasing condition".

Outline of the Proof (Cont)

Most important fact: we can treat $T_1(k)$ and $T_2(k)$ separately, i.e., first consider whether $\tau \in \Theta$, if not, iterate until $\tau \in \Theta$; after this is satisfied, consider the "increasing condition".

Hence, we can argue as follows:

- $T_1(k) < \infty$ (may not be uniform);
- $T_2(k) < A$ uniformly for some A;
- "increasing condition" and strong concavity implies $\{\theta_k\}_{k=k_0}^{\infty}$ in a compact subset of Θ, this in turn will be sufficient for the uniform boundedness of $T_1(k)$.
- Finally, exponential convergence of the algorithm can be obtained.

Complexity of our Algorithm

- When apply our algorithm to compute the channel capacity of finite-state channels with Markovian inputs, the computation complexity of

$$f_k(\theta) = H(X_1^2) + H(Y_k|Y_1^{k-1}) - H(X_k, Y_k|X_1^{k-1}, Y_1^{k-1})$$

  is at most exponential in k. Hence, our algorithm achieves an exponential accuracy in an exponential time. By using change of variable, polynomial accuracy can be achieved within polynomial amount of time.

Applications

- Example 1 (BEC with $(1, \infty)$-RLL input constraint):

$$Y_n = X_n \cdot E_n$$

where X_n binary Markov chain with transition matrix

$$\Pi = \begin{bmatrix} 1-\theta & \theta \\ 1 & 0 \end{bmatrix},$$

and $\{E_n\}$ i.i.d., independent with $\{X_n\}$ and

$$P(E_n = 1) = \varepsilon = 0.1.$$

Applications

In this case,

$$I(X(\theta); Y(\theta)) = (1-\varepsilon)^2 \sum_{j=0}^{\infty} H(X_{j+2}(\theta)|X_1(\theta))\varepsilon^j$$

and is concave with respect to θ (Li, Han, 2014).

Applications

In this case,

$$I(X(\theta); Y(\theta)) = (1-\varepsilon)^2 \sum_{j=0}^{\infty} H(X_{j+2}(\theta)|X_1(\theta))\varepsilon^j$$

and is concave with respect to θ (Li, Han, 2014).

By running our algorithm, we get

$$0.4422382 \leq C \leq 0.4422398.$$

Applications

In this case,

$$I(X(\theta); Y(\theta)) = (1-\varepsilon)^2 \sum_{j=0}^{\infty} H(X_{j+2}(\theta)|X_1(\theta))\varepsilon^j$$

and is concave with respect to θ (Li, Han, 2014).

By running our algorithm, we get

$$0.4422382 \leq C \leq 0.4422398.$$

- By applying our algorithm on the second order Markovian input case, we can show second-order Markov capacity is strictly larger than the first-order Markov capacity.

Applications

- Example 2: A finite-state channel

$$Y_n = \phi(X_n, S_{n-1}), \quad n = 1, 2, \ldots$$

where $\{X_n\}$ is a binary Markov chain, the state $S_n = X_n$ for all n and ϕ is a sliding block code:

$$\phi(00) = 1, \phi(01) = 0, \phi(10) = 0, \phi(11) = 0.$$

Applications

- Example 2: A finite-state channel

$$Y_n = \phi(X_n, S_{n-1}), \quad n = 1, 2, \ldots$$

where $\{X_n\}$ is a binary Markov chain, the state $S_n = X_n$ for all n and ϕ is a sliding block code:

$$\phi(00) = 1, \phi(01) = 0, \phi(10) = 0, \phi(11) = 0.$$

In this case, by "unambiguous formula" for hidden Markov chain, we get:

$$\begin{aligned} I(X;Y) &= \lim_{k\to\infty} H(Y_{k+1}|Y_1^k) \\ &= \sum_{k=1}^{\infty} P(Y_1^k = 1\underbrace{00\ldots00}_{k-1})H(Y_{k+1}|1\underbrace{00\ldots00}_{k-1}). \end{aligned}$$

Suppose $\{(X_n, X_{n-1})\}$ has the transition probability matrix (indexed by $00, 01, 10, 11$):

$$\begin{bmatrix} \theta & 1-\theta & 0 & 0 \\ 0 & 0 & \theta & 1-\theta \\ \theta & 1-\theta & 0 & 0 \\ 0 & 0 & \theta & 1-\theta \end{bmatrix},$$

it can be numerically shown $I(X(\theta), Y(\theta))$ is strongly concave with respect to θ and by going through our algorithm, we have

$$0.4291146 \leq I^{(0)}(X; Y) \leq 0.4294638.$$

Suppose $\{(X_n, X_{n-1})\}$ has the transition probability matrix (indexed by $00, 01, 10, 11$):

$$\begin{bmatrix} \theta & 1-\theta & 0 & 0 \\ 0 & 0 & \theta & 1-\theta \\ \theta & 1-\theta & 0 & 0 \\ 0 & 0 & \theta & 1-\theta \end{bmatrix},$$

it can be numerically shown $I(X(\theta), Y(\theta))$ is strongly concave with respect to θ and by going through our algorithm, we have

$$0.4291146 \leq I^{(0)}(X;Y) \leq 0.4294638.$$

- Again, by comparing it to the birch lower bound for the first-order Markovian input case, we can conclude that the first-order Markov capacity is strictly larger that i.i.d. input case.

Generalization to non-concave case

Our algorithm can be generalized to the case where the target function is non-concave, but extra assumptions are needed:

- There are finitely many stationary points of f and they are away from $\partial\Theta$ (Θ is the domain of the parameter);
- For proper choice of k_0 (large enough), there exists a y_0 such that

$$B := \{x : f_{k_0}(x) \geq y_0\}$$

  is convex, in Θ and contains all the stationary points;
- Choose θ_{k_0} such that $\theta_{k_0} \in B$.

Then we can propose another similar algorithm and prove the local converges.

The second modified gradient descent algorithm.

Step 0. Set $k = 0$, and choose $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$, $\theta_{k_0} \in \Theta$, $k_0 > 0$ and $b \in (0, 1)$ such that

$$\rho^{1/3} + \rho^{2k_0/3} < 1, \quad ||\nabla f_{k_0}(\theta_{k_0})||_2 \geq \frac{2N\rho^{k_0/3}}{1-b}.$$

Step 1. Set $t = 1$ and increase k by 1.

Step 2. Set

$$\tau = \theta_{k-1} + t\nabla f_{k-1}(\theta_{k-1}),$$

If $\tau \notin \Theta$ or

$$||\nabla f_k(\tau)||_2 < \frac{2N\rho^{k/3}}{1-b}$$

or

$$f_k(\tau) < f_k(\theta_{k-1}) + \alpha t||\nabla f_{k-1}(\theta_{k-1})||_2^2,$$

set $t = \beta t$ and go to Step 2, otherwise set $\theta_k = \tau$ and go to Step 1.

Thank You!