Mirror Descent and Blahut-Arimoto Algorithms

K. He[1] J. Saunderson[1] H. Fawzi[2]

[1]Department of Electrical and Computer System Engineering
Monash University

[2]Department of Applied Mathematics and Theoretical Physics
University of Cambridge

SIAM Conference on Optimization, May 2023

Quantum Channel Capacities

Classical-quantum (cq) channel capacity (Schumacher & Westmoreland, 1997), (Holevo, 1998):

$$\max_{p \in \Delta} \quad S\left(\sum_{j=1}^{m} p_j X_j\right) - \sum_{j=1}^{m} p_j S(X_j)$$

where X_j are positive semidefinite matrices and

$$S(X) := -\operatorname{tr}[X \log(X)] \qquad \text{(Quantum entropy)}$$

Objective is concave in p.

Quantum Channel Capacities

Classical-quantum (cq) channel capacity (Schumacher & Westmoreland, 1997), (Holevo, 1998):

$$\max_{p \in \Delta} \quad S\left(\sum_{j=1}^{m} p_j X_j\right) - \sum_{j=1}^{m} p_j S(X_j)$$

where X_j are positive semidefinite matrices and

$$S(X) := -\operatorname{tr}[X \log(X)] \qquad \text{(Quantum entropy)}$$

Objective is concave in p.

How can we efficiently compute this quantity?

Quantum Channel Capacities

Classical-quantum (cq) channel capacity (Schumacher & Westmoreland, 1997), (Holevo, 1998):

$$\max_{p \in \Delta} \quad S\left(\sum_{j=1}^{m} p_j X_j\right) - \sum_{j=1}^{m} p_j S(X_j)$$

where X_j are positive semidefinite matrices and

$$S(X) := -\operatorname{tr}[X \log(X)] \qquad \text{(Quantum entropy)}$$

Objective is concave in p.

How can we efficiently compute this quantity?

- SDP approx. (Fawzi et al., 2019)? But no practical SDP solver for large-scale problems of this type

Quantum Channel Capacities

Classical-quantum (cq) channel capacity (Schumacher & Westmoreland, 1997), (Holevo, 1998):

$$\max_{p \in \Delta} \quad S\left(\sum_{j=1}^{m} p_j X_j\right) - \sum_{j=1}^{m} p_j S(X_j)$$

where X_j are positive semidefinite matrices and

$$S(X) := -\operatorname{tr}[X \log(X)] \qquad \text{(Quantum entropy)}$$

Objective is concave in p.

How can we efficiently compute this quantity?

- SDP approx. (Fawzi et al., 2019)? But no practical SDP solver for large-scale problems of this type
- Gradient-descent-type algorithms don't work well

Table of Contents

Table of Contents

Preliminaries

States:

- Classical: Δ probability distribution
- Quantum: $\mathcal{D}$ density matrix (Hermitian, PSD, unit trace)

Preliminaries

States:

- Classical: Δ probability distribution
- Quantum: $\mathcal{D}$ density matrix (Hermitian, PSD, unit trace)

Let $\log(X) = \sum_i \log(\lambda_i) v_i v_i^\top$ for $X = \sum_i \lambda_i v_i v_i^\top$.

Entropy:

- Classical: $H(x) := -\sum_{i=1}^n x_i \log(x_i)$
- Quantum: $S(X) := -\operatorname{tr}[X \log(X)] = H(\lambda)$

Preliminaries

States:

- Classical: Δ probability distribution
- Quantum: $\mathcal{D}$ density matrix (Hermitian, PSD, unit trace)

Let $\log(X) = \sum_i \log(\lambda_i) v_i v_i^\top$ for $X = \sum_i \lambda_i v_i v_i^\top$.

Entropy:

- Classical: $H(x) := -\sum_{i=1}^n x_i \log(x_i)$
- Quantum: $S(X) := -\operatorname{tr}[X \log(X)] = H(\lambda)$

Relative entropy:

- Classical: $H(x \,\|\, y) := \sum_{i=1}^n x_i \log(x_i / y_i)$
- Quantum: $S(X \,\|\, Y) := \operatorname{tr}[X(\log(X) - \log(Y))]$

Blahut-Arimoto algorithm (Ramakrishnan et al., 2021)

Blahut-Arimoto algorithm first introduced (Blahut, 1972) and (Arimoto, 1972) to solve for classical channel capacities.

Extended to quantum channel capacities in (Nagaoka, 1998), (Li & Cai, 2019), (Ramakrishnan et al., 2021).

Blahut-Arimoto algorithm (Ramakrishnan et al., 2021)

Blahut-Arimoto algorithm first introduced (Blahut, 1972) and (Arimoto, 1972) to solve for classical channel capacities.

Extended to quantum channel capacities in (Nagaoka, 1998), (Li & Cai, 2019), (Ramakrishnan et al., 2021).

$$\min_{x\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(x)\rangle}_{f(x)} = \min_{x\in\mathcal{D}} \min_{y\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(y)\rangle + L\, S(x \,\|\, y)}_{g(x,y)}.$$

for some function $\mathcal{F} : \mathbb{H}^n \to \mathbb{H}^n$ and constant $L > 0$.

Blahut-Arimoto algorithm (Ramakrishnan et al., 2021)

Blahut-Arimoto algorithm first introduced (Blahut, 1972) and (Arimoto, 1972) to solve for classical channel capacities.

Extended to quantum channel capacities in (Nagaoka, 1998), (Li & Cai, 2019), (Ramakrishnan et al., 2021).

$$\min_{x\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(x)\rangle}_{f(x)} = \min_{x\in\mathcal{D}} \min_{y\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(y)\rangle + L\, S(x \,\|\, y)}_{g(x,y)}.$$

for some function $\mathcal{F} : \mathbb{H}^n \to \mathbb{H}^n$ and constant $L > 0$.

$$\text{e.g.} \quad \mathcal{F}(p) = \sum_{j=1}^{m} \boldsymbol{e}_j \boldsymbol{e}_j^\top \operatorname{tr}\left[X_j \left(\log(X_j) - \log\left(\sum_{j=1}^{m} p_j X_j\right)\right)\right].$$

Blahut-Arimoto algorithm (Ramakrishnan et al., 2021)

Blahut-Arimoto algorithm first introduced (Blahut, 1972) and (Arimoto, 1972) to solve for classical channel capacities.

Extended to quantum channel capacities in (Nagaoka, 1998), (Li & Cai, 2019), (Ramakrishnan et al., 2021).

$$\min_{x\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(x)\rangle}_{f(x)} = \min_{x\in\mathcal{D}} \min_{y\in\mathcal{D}} \underbrace{\langle x, \mathcal{F}(y)\rangle + L\, S(x \,\|\, y)}_{g(x,y)}.$$

for some function $\mathcal{F} : \mathbb{H}^n \to \mathbb{H}^n$ and constant $L > 0$.

$$\text{e.g.} \quad \mathcal{F}(p) = \sum_{j=1}^{m} \boldsymbol{e}_j \boldsymbol{e}_j^\top \operatorname{tr}\left[X_j\left(\log(X_j) - \log\left(\sum_{j=1}^{m} p_j X_j\right)\right)\right].$$

Solves by using alternating optimization

$$y^{k+1} = \arg\min_{y\in\mathcal{D}} \, g(x^k, y),$$

$$x^{k+1} = \arg\min_{x\in\mathcal{D}} \, g(x, y^{k+1}).$$

Blahut-Arimoto algorithm (Ramakrishnan et al., 2021)

If $\mathcal{F}$ is continuous and satisfies

$$\mu S(x \| y) \leq \langle x, \mathcal{F}(x) - \mathcal{F}(y) \rangle \leq L S(x \| y),$$

for all $x, y \in \operatorname{relint} \mathcal{D}$ and some $\mu \geq 0$, then

$$\begin{aligned} y^{k+1} &= x^k \\ x^{k+1} &= \frac{\exp(\log(y^{k+1}) - \mathcal{F}(y^{k+1})/L)}{\operatorname{tr}[\exp(\log(y^{k+1}) - \mathcal{F}(y^{k+1})/L)]}. \end{aligned}$$

and BA converges

- sublinearly $O(1/k)$; or
- linearly $O((1 - \mu/L)^k)$ if $\mu > 0$

Blahut-Arimoto algorithm

Theorem 1 (HSF, 2023)

Consider quantum Blahut-Arimoto with continuous $\mathcal{F}$ such that

$$\mu S(x \| y) \leq \langle x, \mathcal{F}(x) - \mathcal{F}(y) \rangle \leq L S(x \| y),$$

The quantum Blahut-Arimoto iterates are equivalent to mirror descent iterates applied to solve

$$\min_{x \in \mathcal{C}} \quad f(x)$$

where

- $\nabla f(x) = \mathcal{F}(x)$ and $f(x) = \langle x, \mathcal{F}(x) \rangle = \langle x, \nabla f(x) \rangle$
- $\mathcal{C} = \mathcal{D}$, kernel function $-S$, step size $t_k = 1/\gamma$,
- f is L-smooth relative to $-S$,
- f is μ-strongly convex relative to $-S$ if $\mu > 0$.

Mirror descent

Consider constrained convex optimization problem

$$\min_{x \in \mathcal{X}} \quad f(x).$$

Projected gradient descent can be represented as

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \langle \nabla f(x^k), x \rangle + \frac{1}{2t_k} \|x - x^k\|_2^2$$

Mirror descent

Consider constrained convex optimization problem

$$\min_{x \in \mathcal{X}} \quad f(x).$$

Projected gradient descent can be represented as

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \langle \nabla f(x^k), x \rangle + \frac{1}{2t_k} \|x - x^k\|_2^2$$

Mirror descent replaces Euclidean norm with Bregman divergence

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \langle \nabla f(x^k), x \rangle + \frac{1}{t_k} D_\varphi(x \| y)$$

where

$$D_\varphi(x \| y) := \varphi(x) - (\varphi(y) + \langle \nabla \varphi(y), x - y \rangle).$$

Mirror descent

Mirror descent w/ $\mathcal{X} = \mathcal{D}$, $\varphi(x) = -S(x)$, $D_\varphi(x \,\|\, y) = S(x \,\|\, y)$

$$x^{k+1} = \operatorname*{arg\,min}_{x \in \mathcal{D}} \langle \nabla f(x^k), x \rangle + \frac{1}{t_k} S(x \,\|\, y)$$
$$= \frac{\exp(\log(x^k) - t_k \nabla f(x^k))}{\operatorname{tr}[\exp(\log(x^k) - t_k \nabla f(x^k))]}.$$

Mirror descent

Mirror descent w/ $\mathcal{X} = \mathcal{D}$, $\varphi(x) = -S(x)$, $D_\varphi(x \,\|\, y) = S(x \,\|\, y)$

$$\begin{aligned} x^{k+1} &= \operatorname*{arg\,min}_{x \in \mathcal{D}} \langle \nabla f(x^k), x \rangle + \frac{1}{t_k} S(x \,\|\, y) \\ &= \frac{\exp(\log(x^k) - t_k \nabla f(x^k))}{\operatorname{tr}[\exp(\log(x^k) - t_k \nabla f(x^k))]}. \end{aligned}$$

Recall Blahut-Arimoto iterate, where $y^{k+1} = x^k$

$$\begin{aligned} x^{k+1} &= \operatorname*{arg\,min}_{x \in \mathcal{D}} \langle \mathcal{F}(x^k), x \rangle + L S(x \,\|\, x^k) \\ &= \frac{\exp(\log(x^k) - \mathcal{F}(x^k)/L)}{\operatorname{tr}[\exp(\log(x^k) - \mathcal{F}(x^k)/L)]}. \end{aligned}$$

Blahut-Arimoto algorithm

Theorem 1 (HSF, 2023)

Consider quantum Blahut-Arimoto with continuous $\mathcal{F}$ such that

$$\mu S(x\|y) \leq \langle x, \mathcal{F}(x) - \mathcal{F}(y) \rangle \leq L S(x\|y),$$

The quantum Blahut-Arimoto iterates are equivalent to mirror descent iterates applied to solve

$$\min_{x \in \mathcal{C}} \quad f(x)$$

where

- $\nabla f(x) = \mathcal{F}(x)$ and $f(x) = \langle x, \mathcal{F}(x) \rangle = \langle x, \nabla f(x) \rangle$
- $\mathcal{C} = \mathcal{D}$, kernel function $-S$, step size $t_k = 1/\gamma$,
- f is L-smooth relative to $-S$,
- f is μ-strongly convex relative to $-S$ if $\mu > 0$.

Relative smoothness

(Bauschke et al., 2017) and (Lu et al., 2018)

A function f is L-smooth relative to φ if for $L > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L(D_\varphi(x \| y) + D_\varphi(y \| x))$$

A function f is μ-strongly convex relative to φ if for $\mu > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu(D_\varphi(x \| y) + D_\varphi(y \| x))$$

Relative smoothness

(Bauschke et al., 2017) and (Lu et al., 2018)

A function f is L-smooth relative to φ if for $L > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L(D_\varphi(x \| y) + D_\varphi(y \| x))$$

A function f is μ-strongly convex relative to φ if for $\mu > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu(D_\varphi(x \| y) + D_\varphi(y \| x))$$

$$f(p) = -S\left(\sum_{j=1}^{m} p_j X_j\right) + \sum_{j=1}^{m} p_j S(X_j)$$

- f is not smooth relative to $\|\cdot\|_2^2/2$ (i.e., gradient not Lipschitz)
- f is 1-smooth relative to $-S(\cdot)$

Quantum Channel Capacities

(Bauschke et al., 2017) and (Lu et al., 2018)

A function f is L-smooth relative to φ if for $L > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L(D_\varphi(x \| y) + D_\varphi(y \| x))$$

A function f is μ-strongly convex relative to φ if for $\mu > 0$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu(D_\varphi(x \| y) + D_\varphi(y \| x))$$

Mirror descent w/ $t_k = 1/L$

- Converges sublinearly $O(1/k)$ if f is L-smooth relative to φ
- Converges linearly $O((1 - \mu/L)^k)$ if f is **also** μ-strongly convex relative to φ

Relative smoothness

Sublinear convergence:

$$\begin{aligned} \text{MD}: &\quad 0 \le \langle x - y, \nabla f(x) - \nabla f(y) \rangle \le L(S(x \| y) + S(y \| x)) \\ \text{BA}: &\quad 0 \le \langle x, \nabla f(x) - \nabla f(y) \rangle \qquad \le LS(x \| y) \end{aligned}$$

Linear convergence:

$$\begin{aligned} \text{MD}: &\quad \mu(S(x \| y) + S(y \| x)) \le \langle x - y, \nabla f(x) - \nabla f(y) \rangle \\ \text{BA}: &\quad \mu S(x \| y) \qquad \le \langle x, \nabla f(x) - \nabla f(y) \rangle \end{aligned}$$

Relative smoothness

Sublinear convergence:

$$\begin{aligned} \text{MD}: &\quad 0 \le \langle x - y, \nabla f(x) - \nabla f(y) \rangle \le L(S(x \| y) + S(y \| x)) \\ \text{BA}: &\quad 0 \le \langle x, \nabla f(x) - \nabla f(y) \rangle \qquad \le LS(x \| y) \end{aligned}$$

Linear convergence:

$$\begin{aligned} \text{MD}: &\quad \mu(S(x \| y) + S(y \| x)) \le \langle x - y, \nabla f(x) - \nabla f(y) \rangle \\ \text{BA}: &\quad \mu S(x \| y) \qquad\qquad \le \langle x, \nabla f(x) - \nabla f(y) \rangle \end{aligned}$$

If $f(x) = \langle x, \nabla f(x) \rangle$, then

$$\underbrace{\begin{aligned} \langle x, \nabla f(x) - \nabla f(y) \rangle \le LS(x \| y) \\ \langle x, \nabla f(x) - \nabla f(y) \rangle \ge \mu S(x \| y) \end{aligned}}_{\text{Conditions for BA convergence}} \quad \begin{aligned} \iff \\ \iff \end{aligned} \quad \underbrace{\begin{aligned} &f \text{ is } L\text{-smooth rel. to } -S \\ &f \text{ is } \mu\text{-strong convex rel. to } -S \end{aligned}}_{\text{Conditions for MD convergence}}$$

Blahut-Arimoto algorithm

Theorem 1 (HSF, 2023)

Consider quantum Blahut-Arimoto with continuous $\mathcal{F}$ such that

$$\mu S(x \| y) \leq \langle x, \mathcal{F}(x) - \mathcal{F}(y) \rangle \leq L S(x \| y),$$

The quantum Blahut-Arimoto iterates are equivalent to mirror descent iterates applied to solve

$$\min_{x \in \mathcal{C}} \quad f(x)$$

where

- $\nabla f(x) = \mathcal{F}(x)$ and $f(x) = \langle x, \mathcal{F}(x) \rangle = \langle x, \nabla f(x) \rangle$
- $\mathcal{C} = \mathcal{D}$, kernel function $-S$, step size $t_k = 1/\gamma$,
- f is L-smooth relative to $-S$,
- f is μ-strongly convex relative to $-S$ if $\mu > 0$.

Table of Contents

Constrained Channel Capacities

Constrained classical-quantum (cq) channel capacity:

$$\begin{aligned} \max_{p\in\Delta} \quad & S\left(\sum_{j=1}^{m} p_j\rho_j\right) - \sum_{j=1}^{m} p_j S(\rho_j) \\ \text{subj. to} \quad & Ap \leq b, \end{aligned}$$

Constrained Channel Capacities

Constrained classical-quantum (cq) channel capacity:

$$\max_{p\in\Delta} \quad S\left(\sum_{j=1}^{m} p_j\rho_j\right) - \sum_{j=1}^{m} p_j S(\rho_j)$$
$$\text{subj. to} \quad Ap \leq b,$$

Blahut-Arimoto algorithms cannot elegantly handle these constraints

$$p^{k+1} \quad = \quad \arg\min_{p\in\Delta} \quad \langle \nabla f(p^k), p\rangle + \frac{1}{t_k} H(p\,\|\,p^k)$$
$$\text{subj. to} \quad Ap \leq b,$$

Primal-dual hybrid gradient

Solve saddle point problem

$$\inf_{x \in \mathcal{C}} \sup_{z \in \mathcal{Z}} \quad \mathcal{L}(x, z) \coloneqq f(x) + \langle z, Ax - b \rangle.$$

Primal-dual hybrid gradient

Solve saddle point problem

$$\inf_{x\in\mathcal{C}} \sup_{z\in\mathcal{Z}} \quad \mathcal{L}(x,z) \coloneqq f(x) + \langle z, Ax - b\rangle.$$

Using primal-dual hybrid gradient

$$\begin{aligned}
\bar{z}^{k+1} &= z^k + \theta_k(z^k - z^{k-1}) \\
x^{k+1} &= \arg\min_{x\in\mathcal{C}} \left\{ \langle \nabla f(x) + A^\dagger \bar{z}^{k+1}, x\rangle + \frac{1}{\tau_k} D_\varphi(x \,\|\, x^k) \right\} \\
z^{k+1} &= \arg\min_{z\in\mathcal{Z}} \left\{ -\langle z, Ax^{k+1} - b\rangle + \frac{1}{2\gamma_k} \|z - z^k\|_2^2 \right\}
\end{aligned}$$

- Ergodic sublinear convergence if f is L-smooth relative to φ.
- Several variations using Bregman divergences e.g., (Chambolle & Pock, 2016), (Jiang & Vandenberghe, 2022)

Constrained Channel Capacities

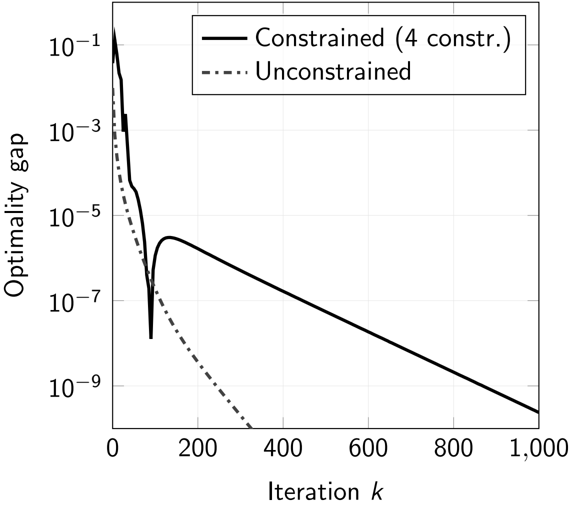
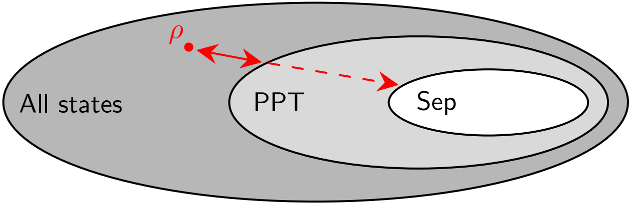


Figure: Classical-quantum channel capacity over $p \in \mathbb{R}^{32}$.



*Entangled states = not separable

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \text{PPT}} \quad S(\rho \| \sigma)$$

where for linear operator $(\cdot)^{T_B}$

$$\text{PPT} = \{\rho \in \mathcal{D} : \rho^{T_B} \succeq 0\},$$

Other applications: Relative entropy of entanglement

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \mathsf{PPT}} \quad S(\rho \| \sigma)$$

$S(\cdot \| \cdot)$ is jointly convex in both arguments.

Other applications: Relative entropy of entanglement

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \text{PPT}} \quad S(\rho \| \sigma)$$

$S(\cdot \| \cdot)$ is jointly convex in both arguments.

Quantum Blahut-Arimoto framework cannot be applied to this problem

Other applications: Relative entropy of entanglement

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \mathrm{PPT}} \quad S(\rho \| \sigma)$$

$S(\cdot \| \cdot)$ is jointly convex in both arguments.

Quantum Blahut-Arimoto framework cannot be applied to this problem

- Objective function does not satisfy $f(x) = \langle x, \nabla f(x) \rangle$

Other applications: Relative entropy of entanglement

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \mathrm{PPT}} \quad S(\rho \| \sigma)$$

$S(\cdot \| \cdot)$ is jointly convex in both arguments.

Quantum Blahut-Arimoto framework cannot be applied to this problem

- Objective function does not satisfy $f(x) = \langle x, \nabla f(x) \rangle$
- Objective function is **not** smooth relative to $-S$

Other applications: Relative entropy of entanglement

(Approximate) relative entropy of entanglement of $\rho \in \mathcal{D}$:

$$\min_{\sigma \in \mathrm{PPT}} \quad S(\rho \| \sigma)$$

$S(\cdot \| \cdot)$ is jointly convex in both arguments.

Quantum Blahut-Arimoto framework cannot be applied to this problem

- Objective function does not satisfy $f(x) = \langle x, \nabla f(x) \rangle$
- Objective function is **not** smooth relative to $-S$

But **is** $\lambda_{max}(\rho)$-smooth and $\lambda_{min}(\rho)$-strongly convex relative to $-\log(\det(\cdot))$.

Other applications: Relative entropy of entanglement

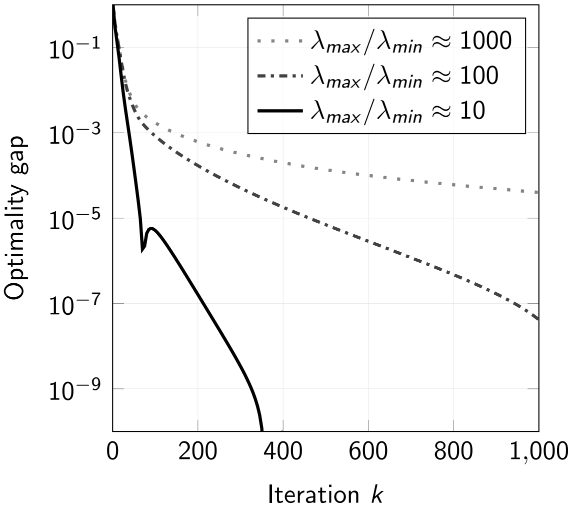


Figure: Relative entropy of entanglement over $\rho \in \mathbb{H}^{25}$.

Conclusion

Summary:

- Blahut-Arimoto algorithms are a specific case of mirror descent and relative smoothness analysis
- Can extend to other applications by using different kernel functions and algorithmic variations of mirror descent

Outlook:

- What other problems in information theory can we extend to?
- Solve general quantum relative entropy programs using similar ideas?

Watch arXiv for incoming preprint!