Q. Inf. Science 3 (8.S372 / 18.S996) — Fall 2020

Assignment 5

Due: **Friday**, *Oct 9, 2020 at* ***5pm*** on canvas.

1. **Entanglement-assisted quantum capacity.**

   (a) *Quantum capacity.* Denote the entanglement-assisted capacity of a quantum channel for sending qubits (resp. cbits) by Q_E (resp. C_E). Relate Q_E to C_E using teleportation and super-dense coding.

   (b) *Classical channels.* The entanglement-assisted capacity theorem states that $C_E(\mathcal{N}) = \max_\tau I(A:B)_\tau$ (see pset 4 for definition of τ). Consider the special case of a classical channel $\mathcal{N}(\rho) := \sum_{x,y} \langle x| \rho |x\rangle N(y|x) |y\rangle\langle y|$ and show that $C_E(\mathcal{N}) = C(N)$. In other words, show that entanglement doesn't increase the capacity of classical channels, and that Shannon's noisy coding theorem can recovered as a special case of the entanglement-assisted capacity theorem.

   (c) *Input concavity.* The formula for C_E from pset 4 can be expressed as

   $$C_E = \max_{\rho_{A'}} C_E(\mathcal{N}, \rho), \tag{1}$$

   where $C_E(\mathcal{N}, \rho) := I(A:B)_\tau$ where $\tau_{AB} = (\mathrm{id}_A \otimes \mathcal{N}_{A' \to B})(\phi^\rho_{AA'})$. Show that $C_E(\mathcal{N}, \rho)$ is independent of the choice of purification ϕ^ρ. Show that $C_E(\mathcal{N}, \rho)$ is concave in the input ρ. [Hint: try purifying $\sum_x p(x) |x\rangle \langle x| \otimes \phi^{\rho_x}$.]

   (d) *Depolarizing channel.* Let $\mathcal{D}^d_p$ (abbreviated $\mathcal{D}$) denote the depolarizing channel on d dimensions with depolarization probability p, defined as

   $$\mathcal{D}^d_p(\rho) = (1-p)\rho + p\frac{I}{d}, \tag{2}$$

   for ρ a d-dimensional density matrix. Observe that $\mathcal{D}(U\rho U^\dagger) = U\mathcal{D}(\rho)U^\dagger$ for any unitary U. Use this property and the input concavity property of C_E to show that $C_E(\mathcal{D}^d_p, \rho)$ is maximized for $\rho = I/d$. Calculate $C_E(\mathcal{D}^d_p, I/d)$.

   (e) *Enhancement from entanglement.* The classical capacity of the depolarizing channel $C(\mathcal{D})$ can be shown to be maximized by applying the HSW theorem to the ensemble where each of the basis states $|1\rangle, \ldots, |d\rangle$ appears with probability $1/d$. Calculate $C(\mathcal{D}^d_p)$. What is the ratio $C_E(\mathcal{D})/C(\mathcal{D})$ in the limits $p \to 0$ and $p \to 1$ as a function of d?

2. **Relative entropy and the matrix multiplicative weight method.**
In this problem we will see how quantum relative entropy can be a useful tool in classical optimization algorithms that have applications in machine learning. For convenience, take log to be base-e in this problem. Some formulas that may be helpful:

$$\ln(A+B) = \ln(A) + \int_0^\infty \mathrm{d}z\,(A+zI)^{-1}B(A+B+zI)^{-1} \tag{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{A(t)} = \int_0^1 \mathrm{d}s\, e^{sA}\frac{\mathrm{d}A}{\mathrm{d}t}e^{(1-s)A} \tag{4}$$

(a) *Variants of gradient descent.* Consider the problem of minimizing a function $f:\mathbb{R}^d \to \mathbb{R}$. We will discuss three algorithms for this problem.

i. *Proximal gradient descent.* The idea of gradient descent is to start with a point x_0 and then in the t^{th} step, move from x_t in the direction of $-\boldsymbol{\nabla} f(x_t)$, i.e. -1 times the gradient of f evaluated at x_t. At the same time, we don't want to move too far from x_t. These goals (moving in the direction of $-\boldsymbol{\nabla} f$ but not too far from x_t) compete and we choose x_{t+1} according to

$$x_{t+1} = \arg\min_{x_{t+1}} \eta\langle x_{t+1}-x_t, \boldsymbol{\nabla} f(x_t)\rangle + \frac{1}{2}\|x_{t+1}-x_t\|_2^2, \tag{5}$$

for some parameter $\eta > 0$. Solve for x_{t+1} in terms of x_t, η, and f. Does this correspond to a step in the direction of $-\boldsymbol{\nabla} f$?

ii. *Mirror descent on probabilities.* Let Δ_d be the set of probability distributions on d items, i.e. $\Delta_d = \{x \in \mathbb{R}^d : \forall i\, x(i) \geq 0, \sum_{i=1}^d x(i) = 1\}$. For probability distributions it is more natural to use the relative entropy as a distance measure instead of the ℓ_2 norm. Thus the *mirror descent* algorithm chooses x_{t+1} according to

$$x_{t+1} = \arg\min_{x_{t+1}} \eta\langle x_{t+1}-x_t, \boldsymbol{\nabla} f(x_t)\rangle + D(x_{t+1}\|x_t). \tag{6}$$

(The terminology "mirror descent" comes from a generalization using something known as as "mirror map" which we will not use in this pset.) Solve for x_{t+1} in terms of x_t, η, and f. As a hint, the update rule you find is called the "multiplicative weights" update rule.

iii. *Mirror descent on density matrices.* Now let $\mathcal{D}_d$ denote $d \times d$ density matrices and define $f : \mathcal{D}_d \mapsto \mathbb{R}$. Note that $\boldsymbol{\nabla} f$ is now a matrix, and for matrices A, B, we define $\langle A, B\rangle := \operatorname{tr}\big[A^\dagger B\big]$. Mirror descent here corresponds to the update rule

$$\rho_{t+1} = \arg\min_{\rho_{t+1}} \eta\langle \rho_{t+1}-\rho_t, \boldsymbol{\nabla} f(\rho_t)\rangle + D(\rho_{t+1}\|\rho_t). \tag{7}$$

Solve for ρ_{t+1} as a function of ρ_t, η and f, assuming for simplicity that ρ_t is full rank. As a hint, you may find the solution of problem 3(b) on pset 2 helpful.

iv. *Continuous-time matrix mirror descent.* It is sometimes more convenient to work with a continuous-time version of the map in (7). Let $\rho(t)$ be a function of t and for $t \geq 0$ let

$$\rho(t+\mathrm{d}t) = \arg \min_{\rho(t+\mathrm{d}t)} \eta \,\mathrm{d}t \,\langle \rho(t+\mathrm{d}t) - \rho(t), \boldsymbol{\nabla} f(\rho(t))\rangle + D(\rho(t+\mathrm{d}t) \| \rho(t)). \quad (8)$$

Write down a differential equation for $\ln \rho(t)$. [Hint: instead of solving (8) directly, guess the form of the answer by analogy with your answer from part iii.]

(b) *Convergence of matrix multiplicative weights.* Let ρ_* be an arbitrary density matrix (which we will later take to be the minimizer of f).

i. *Progress.* Using the above differential equation show that

$$\frac{\mathrm{d}}{\mathrm{d}t} D(\rho_* \| \rho(t)) = \langle \eta \boldsymbol{\nabla} f(\rho(t)), \rho_* - \rho(t)\rangle \quad (9)$$

ii. *Convexity.* Suppose that f is convex. Prove that

$$f(\sigma_1) - f(\sigma_2) \leq \langle \boldsymbol{\nabla} f(\sigma_1), \sigma_1 - \sigma_2\rangle, \quad (10)$$

for any density matrices σ_1, σ_2.

iii. *Convergence.* Let $\rho(0) = I/d$ and let $\rho_* = \arg\min f(\rho_*)$.
Show that $D(\rho_* \| \rho(0)) \leq \log(d)$. How large should T be to guarantee that $f(\rho(T)) \leq f(\rho_*) + \epsilon$? You may find it helpful to show that $\frac{df(\rho(t))}{dt} \leq 0$, which can be done either using the fact that $\rho(t)$ optimizes (8) or with direct calculation.
Observe that relative entropy is used in the analysis but the final bound is only in terms of f and the update rule you derived can also be expressed without referencing the relative entropy. Your solution turns out to slightly overstate the power of this algorithm since actual computers need to work in discrete time and this introduces some additional difficulties. Still the mirror descent algorithm is a very powerful tool because of its favorable dependence on d.