Problem 1 FEL 3310 Distributed Optimization

Solution

(a) 10 $\text{Null}(\mathcal{L}_G) = \{ \mathbf{z} = [z_1, \cdots, z_N] \in \mathbb{R}^N : z_i = z_j \; i,j = 1, \cdots, N \}$

Thus, $\mathcal{L}_G \mathbf{x} = 0 \iff x_i = x_j \quad \forall i,j \in \mathcal{V}$.

$\mathbf{x}^* = [x^*, \cdots, x^*] \in \mathbb{R}^N$, where x^* is the optimal solution to problem (1).

(b) 10 For any $\beta \in (0,1)$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$.

$$
\begin{aligned}
f(\beta \mathbf{x} + (1-\beta)\mathbf{y}) &= \sum_{i \in \mathcal{V}} f_i(\beta x_i + (1-\beta) y_i) \\
&\leq \sum_{i \in \mathcal{V}} \beta f_i(x_i) + (1-\beta) f_i(y_i) - \frac{\mu_i}{2}\beta(1-\beta)\|x_i - y_i\|^2 \\
&\leq \beta \sum_{i \in \mathcal{V}} f_i(x_i) + (1-\beta) \sum_{i \in \mathcal{V}} f_i(y_i) - \frac{\mu}{2}\beta(1-\beta)\sum_{i \in \mathcal{V}} \|x_i - y_i\|^2 \\
&= \beta f(\mathbf{x}) + (1-\beta) f(\mathbf{y}) - \frac{\mu}{2}\beta(1-\beta)\|\mathbf{x} - \mathbf{y}\|^2.
\end{aligned}
$$

(c) 10

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) &= \sum_{i \in \mathcal{V}} f_i(x_i) + \lambda_i \Big(\sum_{j \in \mathcal{N}_i} (x_i - x_j) \Big) \\
&= \sum_{i \in \mathcal{V}} f_i(x_i) + \sum_{j \in \mathcal{N}_i} (\lambda_i - \lambda_j) x_i
\end{aligned}
$$

(d) The dual function

20

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathbf{X}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x}) + \boldsymbol{\lambda}^T (\mathcal{L}_G \mathbf{x})$$

where $\mathbf{X} = \{ \mathbf{z} = [z_1, \cdots, z_N]^T \in \mathbb{R}^N : z_i \in X_i \ \ \forall i = 1, \cdots, N \}$.

Let $\mathbf{x}^*(\boldsymbol{\lambda}) = \arg\min_{\mathbf{x} \in \mathbf{X}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.

Note that for any $\boldsymbol{\lambda} \in \mathbb{R}^N$, $\mathbf{x}^*(\boldsymbol{\lambda})$ uniquely exists due to the strong convexity of $f(\mathbf{x})$ proved in (b).

Therefore, $g(\boldsymbol{\lambda})$ is differentiable and

$$\nabla g(\boldsymbol{\lambda}) = \mathcal{L}_G \mathbf{x}^*(\boldsymbol{\lambda}).$$

For any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathbb{R}^N$,

$$\| \nabla g(\boldsymbol{\lambda}_1) - \nabla g(\boldsymbol{\lambda}_2) \|^2 = \| \mathcal{L}_G (\mathbf{x}^*(\boldsymbol{\lambda}_1) - \mathbf{x}^*(\boldsymbol{\lambda}_2)) \|^2$$

$$\leq \lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G) \, \| \mathbf{x}^*(\boldsymbol{\lambda}_1) - \mathbf{x}^*(\boldsymbol{\lambda}_2) \|^2. \quad \#$$

From the first-order optimality condition, there exist subgradients $s(\mathbf{x}^*(\boldsymbol{\lambda}_1)) \in \partial f(\mathbf{x}^*(\boldsymbol{\lambda}_1))$ and $s(\mathbf{x}^*(\boldsymbol{\lambda}_2)) \in \partial f(\mathbf{x}^*(\boldsymbol{\lambda}_2))$

s.t.

$$(s(\mathbf{x}^*(\boldsymbol{\lambda}_1)) + \mathcal{L}_G \boldsymbol{\lambda}_1)^T (\mathbf{x}^*(\boldsymbol{\lambda}_2) - \mathbf{x}^*(\boldsymbol{\lambda}_1)) \geq 0$$

$$(s(\mathbf{x}^*(\boldsymbol{\lambda}_2)) + \mathcal{L}_G \boldsymbol{\lambda}_2)^T (\mathbf{x}^*(\boldsymbol{\lambda}_1) - \mathbf{x}^*(\boldsymbol{\lambda}_2)) \geq 0$$

Adding these two inequalities, we have

$$\left(s(x^*(\lambda_1)) - s(x^*(\lambda_2)) \right)^T \left(x^*(\lambda_1) - x^*(\lambda_2) \right)$$

$$\leq (\lambda_1 - \lambda_2)^T \mathcal{L}_G \left(x^*(\lambda_2) - x^*(\lambda_1) \right)$$

Since f is strongly convex with convexity parameter μ, the left-hand side of the above inequality is bounded from below by $\mu \| x^*(\lambda_1) - x^*(\lambda_2) \|^2$.

Therefore, $$\| x^*(\lambda_1) - x^*(\lambda_2) \|^2 \leq \frac{\sqrt{\lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G)}\, \|\lambda_1 - \lambda_2\| \cdot \|x^*(\lambda_1) - x^*(\lambda_2)\|}{\mu},$$

i.e., $$\| x^*(\lambda_1) - x^*(\lambda_2) \| \leq \frac{\sqrt{\lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G)}}{\mu} \| \lambda_1 - \lambda_2 \|.$$

This, along with #, gives

$$\| \nabla g(\lambda_1) - \nabla g(\lambda_2) \| \leq \frac{\lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G)}{\mu} \| \lambda_1 - \lambda_2 \|.$$

(e) Let each node $i \in V$ maintains two state variables
$\lambda_i(k) \in \mathbb{R}$ and $x_i(k) \in \mathbb{R}$.

At time $k=0$,

(1) Each node i initializes $\lambda_i(0) \in \mathbb{R}$.

(2) All the nodes agree on some step-size $\alpha > 0$.

At each iteration $k \geq 1$:

(1) Each node i transmits $\lambda_i(k-1)$ to its neighbors.

(2) Upon receiving $\lambda_j(k-1)$ $\forall j \in N_i$, node i updates:

$$x_i(k) = \arg\min_{x_i \in X_i} f_i(x_i) + \sum_{j \in N_i} (\lambda_i(k-1) - \lambda_j(k-1)) x_i$$

(3) Each node i transmits $x_i(k)$ to its neighbors.

(4) Upon receiving $x_j(k)$ $\forall j \in N_i$, node i updates.

$$\lambda_i(k) = \lambda_i(k-1) + \alpha \left(\sum_{j \in N_i} (x_i(k) - x_j(k)) \right).$$

- Because this distributed algorithm is just the gradient method

$$\lambda(k) = \lambda(k-1) + \alpha \nabla g(\lambda(k-1))$$

$$(\lambda(k) = [\lambda_1(k), \cdots, \lambda_N(k)]^T \in \mathbb{R}^N)$$

applied to solve the dual problem of (2), and because ∇g is Lipschitz continuous with Lipschitz constant $\lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G)/\mu$, we have (from Lecture 2)

$$g^* - g(\lambda(k)) \leq \frac{2(g^* - g(\lambda(0))\,\|\lambda(0) - \lambda^*\|^2}{2\|\lambda(0) - \lambda^*\|^2 + (g^* - g(\lambda(0))\,\alpha(2 - L\alpha)k}$$

for $\alpha \in (0,\ 2\mu/\lambda_{max}(\mathcal{L}_G^T \mathcal{L}_G))$,

where g^* is the dual optimal value and $\lambda^* \in \mathbb{R}^N$ is some dual optimal solution.

Since $\|x^*(\lambda) - x^*\| \leq \sqrt{\dfrac{g^* - g(\lambda)}{\mu}}$ (from Lecture 4), we have

$$\|x^*(\lambda(k)) - x^*\| \leq \frac{a}{\sqrt{b + ck}},$$

where
$$a = \sqrt{2(g^* - g(\lambda(0)))\,\|\lambda(0) - \lambda^*\|^2 / \mu}$$
$$b = 2\|\lambda(0) - \lambda^*\|$$
$$c = (g^* - g(\lambda(0)))\,\alpha(2 - L\alpha).$$

Moreover, note that

$$x^*(\lambda(k)) = [x_1(k), \cdots, x_N(k)]^T$$

Hence,

$$\|x^*(\lambda(k)) - x^*\|^2 = \sum_{i \in V} \|x_i(k) - x^*\|^2$$
$$\geq \|x_i(k) - x^*\|^2 \quad \forall i \in V.$$

Consequently,

$$\|x_i(k) - x^*\| \leq \frac{a}{\sqrt{b + ck}} \quad \forall i \in V.$$

(f) $\sum_{i \in V} f_i(x) = 5x^2 - 24x + 4$

20

$$\bigcap_{i \in V} X_i = [-2, 2]$$

$$x^* = 2$$

FEL3310 - Distributed Optimization.

Take-home exam.

Problem 2.

1. a) The initial condition has two satellite nodes with different colors. Since we are selecting from a set of two colors only, a proper coloring will be obtained as soon as the set of satellite nodes ~~are~~ is selected.

Let $S_2 = \{1, \dots, N\}$ and $S_1 = \{0\}$.

At any iteration, $\mathbb{P}[S_2] = \mathbb{P}[S_1] = \frac{1}{2}$

Hence $\mathbb{P}[\tau = k] = \underbrace{\left(\frac{1}{2}\right)^{k-1}}_{\substack{\text{Probability} \\ \text{that for the} \\ \text{first } k-1 \\ \text{iterations,} \\ S_1 \text{ is selected}}} \underbrace{\frac{1}{2}}_{\substack{\text{Probability that} \\ S_2 \text{ is selected}}} = \frac{1}{2^k}$

2. a) Note that the initial coloring ϕ_0 is not defined. It could well be that ϕ_0 is not a proper coloring!

Also observe that $(\phi_n)_{n \geq 1}$ is a Markov chain with transition probabilities given by:

$$f(\phi, \phi_{n+1}) \triangleq \mathbb{P}\left[\phi_{n+1} \mid \phi_n = \phi\right] = \begin{cases} \dfrac{1}{N+1} \dfrac{\lambda^{r_{ij}}}{\sum_{c \in A_\phi(i)} \lambda^{r_{ic}}} & \text{if } \phi_{n+1}(i) = j,\ \phi_{n+1}(i') = \phi(i')\ \forall i' \neq i \\ & \text{and } j \in A_\phi(i) \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that the set of configurations ϕ that are not proper colorings is transient (*) (after a finite time, ϕ_n is a proper coloring).

Let Φ be the set of proper colorings. The Markov chain $(\phi_n)_{n \geq 1}$ is irreducible when starting from a proper coloring: Any proper coloring can be reached with positive probability after a finite time. As a consequence, all proper colorings have the same period. It is straightforward to see that all proper colorings are aperiodic $\left(\mathbb{P}(\phi_{n+1} = \phi \mid \phi_n = \phi) > 0 \right)$.

Thus the Markov chain has a unique stationary distribution π, and $\lim_{n \to \infty} \mathbb{P}\left[\phi_n = \phi \right] = \pi(\phi)$, $\forall \phi \in \Phi$.

(*) Please look at the very nice and concise book by Olle Häggström: "Finite Markov Chains and Algorithmic Applications", Cambridge Univ. Press.

The stationary distribution satisfies the balance equations:

$$\forall \phi \in \Phi, \quad \pi(\phi) = \sum_{\phi' \in \Phi} \pi(\phi') \, f(\phi', \phi) \qquad (1)$$

One can directly verify that the distribution π given below satisfies (1):

$$\forall \phi \in \Phi, \quad \pi(\phi) \propto \lambda^{\sum_i \mu_i \phi(i)}.$$

Alternatively, one may also notice that the Markov chain is reversible in the sense that the above π satisfies the following ~~eqns~~ detailed balance equations :

$$\pi(\phi)\, f(\phi, \phi') = \pi(\phi')\, f(\phi', \phi) \quad , \ \forall\, \phi, \phi' \in \Phi \ .$$

Now when λ is large, π concentrates on proper colorings with maximum social welfares .

b) .

FEL3310 Distributed Optimization.

Take-home exam.

Problem 3.

1 a) We have $\theta_{i,n} = \frac{1}{n} \sum_{t=1}^{n} r_{i,t} \mathbb{1}\{i_t = i\}$

where $i_{t+1} \in \underset{i}{\text{argmax}} \dfrac{r_{i,t+1}}{d + \theta_{i,t}}$.

We simply deduce that:

$$\theta_{i,n+1} = \theta_{i,n} + \frac{1}{n+1}\left[r_{i,n+1} \mathbb{1}_{i_{n+1}=i} - \theta_{i,n} \right]$$

This iteration is that of a stochastic approximation algorithm. (cf the course lecture notes).

Let $\mathcal{F}_n$ be the σ-algebra generated by $(\theta_0, r_0, \dots, \theta_n, r_n)$ where $\theta_t = (\theta_{1,t}, \theta_{2,t})$ and $r_t = (r_{1,t}, r_{2,t})$.

We compute $\mathbb{E}\left[r_{1,n+1} \mathbb{1}_{i_{n+1}=1} \middle| \mathcal{F}_n \right]$:

Remember that $r_{1,n+1} = \frac{1}{\beta_1} \xi_{1,n+1}$ where $\xi_{1,n+1}$ is a r.v. with exponential distribution of unit mean, i.e. it has $\exp(-u)\,du$.

$$\mathbb{E}\left[r_{1,n+1} \mathbb{1}_{\{i_{n+1}=1\}} \,\middle|\, \mathcal{F}_n \right] = \mathbb{E}\left[\mathbb{E}\left[r_{1,n+1} \mathbb{1}_{\{i_{n+1}=1\}} \,\middle|\, r_{1,n+1} \right] \,\middle|\, \mathcal{F}_n \right]$$

$$= \int_0^\infty \frac{u}{\beta_1} \underbrace{\mathbb{P}\left[i_{n+1} = 1 \,\middle|\, \mathcal{F}_n, \xi_{1,n+1} = u \right]}_{A} e^{-u}\, du .$$

Now we have:

$$A = \mathbb{P}\left[\frac{r_{2,n+1}}{d+\theta_{2,n}} < \frac{u}{\beta_1 (d+\theta_{1,n})} \,\middle|\, \mathcal{F}_n \right]$$

$$= \mathbb{P}\left[\xi_{2,n+1} < \frac{d+\theta_{2,n}}{d+\theta_{1,n}} \frac{\beta_2}{\beta_1} u \right]$$

$$= 1 - \exp\left[- \frac{d+\theta_{2,n}}{d+\theta_{1,n}} \cdot \frac{\beta_2}{\beta_1} u \right]$$

Hence: after integration,

$$\mathbb{E}\left[r_{1,n+1} \mathbb{1}_{\{i_{n+1}=1\}} \,\middle|\, \mathcal{F}_n \right] = \frac{1}{\beta_1} - \frac{\beta_1 (d+\theta_{1,n})^2}{\left(\beta_1 (d+\theta_{1,n})^2 + \beta_2 (d+\theta_{2,n})^2 \right)}$$

The same result holds for $i=2$ by symmetry.

We have:

$$\theta_{1,n+1} = \theta_{1,n} + \frac{1}{n+1} \left(f_1(\theta_n) - \theta_{1,n} \right) + \frac{1}{n+1} M_{1,n}$$

where $M_{1,n}$ is a zero-mean random noise and

$$f_1(\theta) = \frac{1}{\beta_1} - \frac{\beta_1 (d+\theta_1)^2}{\left(\beta_1 (d+\theta_1) + \beta_2 (d+\theta_2) \right)^2}$$

$f = (f_1, f_2)$ is Lipschitz continuous, and one can check that Assumptions (A1) - (A4) hold (those in the lecture notes).

Note Assumption (A5) is not required to prove that the algorithm behaves as the ODE:

$$\dot{\theta}_i = f_i(\theta) - \theta_i, \quad i = 1, 2.$$

See Lemma 1 in the lecture notes for a precise statement.

b) Refer to Hirsh's papers.

c) When $d = 0$, the equilibrium points of the ODE satisfy: $\theta_i = f_i(\theta)$, $i = 1, 2$.

At equilibrium, one can easily observe that

$$\beta_1 \theta_1 = \beta_2 \theta_2$$

$$\Rightarrow \quad \theta_i = \frac{3}{4} \frac{1}{\beta_i}, \quad i = 1, 2.$$

Extension to N users: using similar arguments, we get: at equilibrium

$$\theta_i = \frac{1}{\beta_i} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{k}, \quad \forall i.$$

Observe that a gain roughly equal to $\log N$ is achieved when being opportunistic, compared to an oblivious algorithm that would randomly selects users.

2.

Let $U(\theta) = \sum_i \log(d + \theta_i)$ and $\varepsilon_n = \frac{1}{n}$.

We have after Taylor expansion:

$$U(\theta_{n+1}) - U(\theta_n) = \varepsilon_n \underbrace{\sum_i \frac{r_{i,n+1} \mathbb{1}_{i_{n+1}=i} - \theta_{i,n}}{d + \theta_{i,n}}}_{} + O(\varepsilon_n^2)$$

This term is maximized when $i_{n+1} \in \arg\max_i \left\{ \frac{r_{i,n+1}}{d + \theta_{i,n}} \right\}$

The algorithm is, in the limit, a "steepest ascent" algorithm for strictly concave utility functions.

To prove it formally (beyond this heuristic argument) requires a little more. Please look at proof of Theorem 2.3 in: H. Kushner and P. Whiting. "Convergence of Prop: Fair Sharing Algorithms Under General Conditions" IEEE Trans. on Wireless Communications, Vol 3 No 4, July 2004

3) As before we compute:

$$\mathbb{E}\left[r_{1,n+1} \mathbb{1}_{i_{n+1}=1} \middle| \mathcal{F}_n \right] = \int_0^\infty \frac{u}{\beta_1} \underbrace{\mathbb{P}\left[i_{n+1}=1 \middle| \mathcal{F}_n, \xi_{1,n+1}=u \right]}_{A} e^{-u} du$$

Here $A = \mathbb{P}\left[u > \frac{\beta_1}{\beta_2} \frac{d+\theta_{1,n}}{d+\theta_{2,n}} \right] = \mathbb{1}_{u > \frac{\beta_1}{\beta_2} \frac{d+\theta_{1,n}}{d+\theta_{2,n}}}$

$$\Rightarrow \mathbb{E}\left[r_{1,n+1} \mathbb{1}_{i_{n+1}=1} \middle| \mathcal{F}_n \right] = \int_{\frac{\beta_1}{\beta_2} \frac{d+\theta_{1,n}}{d+\theta_{2,n}}}^\infty \frac{u}{\beta_1} e^{-u} du$$

$$= \frac{1}{\beta_1} \left(1 + \frac{\beta_1 (d+\theta_{1,n})}{\beta_2 (d+\theta_{2,n})} \right) \exp\left(- \frac{\beta_1 (d+\theta_{1,n})}{\beta_2 (d+\theta_{2,n})} \right)$$

Similarly

$$\mathbb{E}\left[r_{2,n+1} \mathbb{1}_{i_{n+1}=2} \middle| \mathcal{F}_n \right] = \mathbb{P}\left[\frac{1}{\beta_2 (d+\theta_{2,n})} > \frac{\xi_{1,n+1}}{\beta_1 (d+\theta_{1,n})} \middle| \mathcal{F}_n \right] \frac{1}{\beta_2}$$

$$= \frac{1}{\beta_2} \left(1 - \exp\left(- \frac{\beta_1 (d+\theta_{1,n})}{\beta_2 (d+\theta_{2,n})} \right) \right)$$

The limiting ODE is:

$$\begin{cases} \dot{\theta}_1 = \frac{1}{\beta_1} \left(1 + \frac{\beta_1 (d+\theta_1)}{\beta_2 (d+\theta_2)} \right) \exp\left(- \frac{\beta_1 (d+\theta_1)}{\beta_2 (d+\theta_2)} \right) - \theta_1 \\ \dot{\theta}_2 = \frac{1}{\beta_2} \left(1 - \exp\left(- \frac{\beta_1}{\beta_2} \frac{d+\theta_1}{d+\theta_2} \right) \right) - \theta_2 . \end{cases}$$

For $\alpha = 0$, an equilibrium point would satisfy :

$$\begin{cases} \beta_1 \theta_1 = \left(1 + \dfrac{\beta_1 \theta_1}{\beta_2 \theta_2}\right) \exp\left(- \dfrac{\beta_1 \theta_1}{\beta_2 \theta_2}\right) \\ \beta_2 \theta_2 = 1 - \exp\left(- \dfrac{\beta_1 \theta_1}{\beta_2 \theta_2}\right) \end{cases}$$

let $x = \dfrac{\beta_1 \theta_1}{\beta_2 \theta_2}$, then $x = g(x)$ where $g(x) = (1+x) \dfrac{e^{-x}}{1 - e^{-x}}$

for $x \geqslant 0$, g is positive and decreasing. Hence

$x = g(x)$ has a unique solution greater than 1

$x \simeq 1{,}083$.

$\Rightarrow \theta_1 \simeq \dfrac{0{,}7}{\beta_1}$

$\theta_2 \simeq \dfrac{0{,}66}{\beta_2}$

As a result, the scheduler is not fair in time ...