Adaptive Bound Optimization for Online Convex Optimization

McMahan, H. B., & Streeter, M. (2010)

arXiv

Presenter: Sarah Kim

2019.05.31

1. Introduction

- Consider online convex optimization.
- $\mathcal{F} \subseteq \mathbb{R}^n$: closed, bounded, convex feasible set
- On each round $t = 1, \ldots, T$, pick a point $x_t \in \mathcal{F}$.

  For a given convex loss function f_t,

$$\text{Regret} := \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{F}} \sum_{t=1}^{T} f_t(x).$$

1. Introduction

- Online gradient descent algorithm acheive upper bound of

$$\mathcal{O}(DM\sqrt{T}),$$

  where

  - D: the L_2 diameter of $\mathcal{F}$;
  - M: a bound on L_2 norms of gradients of the loss functions.

- This is minimax optimal when $\mathcal{F}$ is a hypersphere, but we will prove that much better algorithms exist when $\mathcal{F}$ is the hypercube.

1. Introduction

- Hence, we introduce additional parameter $\theta_1, \ldots, \theta_T$ that capture more of the problem's structure.
- We choose θ_t adaptively based on $f_1, \ldots, f_{t-1}$, for $t = 1, \ldots, T$.
- Construct functional upper bounds on reget $B_R(\theta_1, \ldots, \theta_T; f_1, \ldots, f_T)$.
- If for all possible $(f_1, \ldots, f_T)$ we have

$$B_R(\theta_1, \ldots, \theta_T; f_1, \ldots, f_T) \leq \kappa \inf_{\theta'_1, \ldots, \theta'_T \in \Theta^T} B_R(\theta'_1, \ldots, \theta'_T; f_1, \ldots, f_T),$$

then we say the adaptive scheme is κ-**competitive** for the bound optimization problem.

1. Introduction

- **FTPRL (Follow the proximally-regularized leader) algorithm**:

  On round $t+1$, selects

$$x_{t+1} = \underset{x \in \mathcal{F}}{\operatorname{argmin}} \left(\sum_{\tau=1}^{t} (r_\tau(x) + f_\tau(x)) \right),$$

  where

  - $x_1 = 0$ (W.L.O.G., we assume $0 \in \mathcal{F}$)
  - $r_t(x)$: regularization function; $f_t(x)$: convex loss function.
  - Consider regularization functions of the form

$$r_t(x) = \frac{1}{2} \| Q_t^{\frac{1}{2}} (x - x_t) \|_2^2$$

    where Q_t is a positive semidefinite matrix (which is adaptively selected).

Overview

- Notations
  - $\vec{Q_T} = (Q_1, \ldots, Q_T)$;
  - $\vec{g_T} = (g_1, \ldots, g_T)$, where g_t is a subgradient of f_t at x_t;
  - $Q_{1:t} = \sum_{\tau=1}^{t} Q_\tau$
- For a convex set $\mathcal{F}$, define $\mathcal{F}_{\mathrm{sym}} = \{x - x' | x, x' \in \mathcal{F}\}$.

1. Regret bound:

$$\text{Regret} \leq B_R(\vec{Q_T}, \vec{g_T}) := \frac{1}{2} \sum_{t=1}^{T} \max_{\hat{y} \in \mathcal{F}_{\mathrm{sym}}} (\hat{y}^\top Q_t \hat{y}) + \sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t$$

2. We prove competitive ratios w.r.t. B_R for several adaptive schemes for selecting Q_t matrices.
3. Find a fundamental connection between the shape of the feasible set and the importance of choosing the regularization matrices adaptively.

Notations and technical background

- Notations
  - $\partial f(x)$: the set of subgradients of f evaluated at x
  - S^n_+: the set of symm. positive semidefinite $n \times n$ matrices;
    S^n_{++}: the set of symm. positive definite $n \times n$ matrices
  - $\|\cdot\|$: L_2 norm
- Since f_t is convex loss function,

$$f_t(x) \geq g_t^\top (x - x_t) + f_t(x_t),$$

where $g_t \in \partial f(x_t)$. And the above inequality is tight for $x = x_t$. Hence the update of FTPRL is

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{F}} \left(\frac{1}{2} \sum_{\tau=1}^{t} (x - x_\tau)^\top Q_\tau (x - x_\tau) + g_{1:t} \cdot x \right) \qquad (1)$$

2. Analysis of FTPRL

- In this section, we prove the following bound on the regret of FTPRL for an arbitrary seq. of regularization matrices Q_t.

Theorem 2 Let $\mathcal{F} \subseteq \mathbb{R}^n$ be a closed, bounded convex set with $0 \in \mathcal{F}$. Let $Q_1 \in S^n_{++}$, and $Q_2, \ldots, Q_T \in S^n_{+}$. Define $r_t(x) = \frac{1}{2}\|Q_t^{\frac{1}{2}}(x - x_t)\|_2^2$, and $A_t = (Q_{1:t})^{\frac{1}{2}}$. Let f_t be a seq. of loss functions with $g_t \in \partial f_t(x_t)$ a sub-gradient of f_t at x_t. Then the FTPRL algorithm with $x_1 = 0$, and Eq. (1) has a regret bound

$$\text{Regret} \leq r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} \|A_t^{-1} g_t\|^2$$

where $\mathring{x} = \operatorname{argmin}_{x \in \mathcal{F}} f_{1:T}(x)$ is the post-hoc optimal feasible point.

2. Analysis of FTPRL

Proof of Theorem 2

1 First we show that for a seq. of non-negetive functions $r_1, \dots, r_T$,

$$\text{Regret} \le r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} (f_t(x_t) - f_t(x_{t+1}))$$

($\because$) Define $f'_t(x) = f_t(x) + r_t(x)$ and $\hat{x}_t = \text{argmin}_{x \in \mathcal{F}} f'_{1:t}(x)$. Then we have

$$\sum_{t=1}^{T} f'_t(\hat{x}_t) \le \min_{x \in \mathcal{F}} f'_{1:T}(x) \le f'_{1:T}(\mathring{x})$$
$$\Leftrightarrow \sum_{t=1}^{T} f_t(\hat{x}_t) + r_t(\hat{x}_t) \le r_{1:T}(\mathring{x}) + f_{1:T}(\mathring{x})$$

Since $r_t(\hat{x}_t)$ is non-negative, we have

$$\sum_{t=1}^{T} f_t(x_t) - f_{1:T}(\mathring{x}) \le r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} (f_t(x_t) - f_t(x_{t+1}))$$

Proof of Theorem 2 (cont'd)

2 We show that $f_t(x_t) - f_t(x_{t+1}) \le g_t(x_t - x_{t+1}) \overset{?}{\le} \|A_t^{-1} g_t\|^2$.

- (Key idea 1) Let $Q \in S^n_{++}$ and $h \in \mathbb{R}^n$, consider the function

$$f(x) = h^\top x + \frac{1}{2} x^\top Q x.$$

Let $\mathring{u} = \operatorname{argmin}_{u \in \mathbb{R}^n} f(u)$. Then, letting $A = Q^{\frac{1}{2}}$, we have

$$\underset{x \in \mathcal{F}}{\operatorname{argmin}}\, f(x) = \underset{x \in \mathcal{F}}{\operatorname{argmin}}\, \|A(x - \mathring{u})\|.$$

- (Key idea 2) Let $v, g \in \mathbb{R}^n$ and let $u_1 = -Q^{-1} v$ and $u_2 = -Q^{-1}(v + g)$. Then letting $x_1 = \operatorname{argmin}_{x \in \mathcal{F}} \|A(x - u_1)\|$ and $x_2 = \operatorname{argmin}_{x \in \mathcal{F}} \|A(x - u_2)\|$,

$$g^\top (x_1 - x_2) \le \|A^{-1} g\|^2.$$

3. Specific Adaptive Algorithms and Competitive Ratios

- By Thm 2, we have

$$\text{Regret} \leq r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} \|A_t^{-1} g_t\|^2$$
$$\leq \frac{1}{2} \sum_{t=1}^{T} \max_{\hat{y} \in \mathcal{F}_{\text{sym}}} (\hat{y}^\top Q_t \hat{y}) + \sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t =: B_R(\vec{Q}_T, \vec{g_T})$$

- Best post-hoc bound: $\inf_{\vec{Q}_T \in \mathcal{Q}^T} B_R(\vec{Q}_T, \vec{g_T})$, where $\mathcal{Q} \subseteq S_+^n$
- Using the fact that $Q_1, \ldots, Q_T$ are positive semidefinite matrices, one can show that the best post-hoc bound can solve an optimization of the form,

$$\inf_{Q \in \mathcal{Q}} \left(\max_{\hat{y} \in \mathcal{F}_{\text{sym}}} \left(\frac{1}{2} \hat{y}^\top Q \hat{y} \right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \right). \tag{2}$$

3.1. Adaptive coordinate-constant regularization

- We derive bounds where Q_t is chosen from the set $\mathcal{Q}_{\text{const}} := \{\alpha I | \alpha \geq 0\}$.

Corollary 8 Suppose $\mathcal{F}$ has L_2 diameter D. Then, if we run FTPRL with diagonal matrices s.t.

$$(Q_{1:t})_{ii} = \bar{\alpha}_t = \frac{2\sqrt{G_t}}{D}$$

where $G_t = \sum_{\tau=1}^{t} \sum_{i=1}^{n} g_{\tau,i}^2$, then

$$\text{Regret} \leq 2D\sqrt{G_T}.$$

- If $\|g_t\|_2 \leq M$, then $G_T \leq M^2 T$, and this translates to a bound of $\mathcal{O}(DM\sqrt{T})$.
- When $\mathcal{F} = \{x | \|x\|_2 \leq D/2\}$, this bound is $\sqrt{2}$-competitive for the bound optimization problem over $\mathcal{Q}_{\text{const}}$.

Proof of Corollary 8

- Let the diagonal entries of Q_t all be $\alpha_t = \bar{\alpha}_t - \bar{\alpha}_{t-1}$ with $\bar{\alpha}_0$, then $\alpha_{1:t} = \bar{\alpha}_t$. Note $\alpha_t \geq 0$, so this choice is feasible.
- Left of $B_R(\vec{Q}_T, \vec{g_T})$:

  letting $\hat{y}_t$ be an arbitrary seq. of points from $\mathcal{F}_{\text{sym}}$, and noting $\hat{y}_t^\top \hat{y}_t \leq D^2$,

$$\frac{1}{2}\sum_{t=1}^{T} \hat{y}_t^\top Q_t \hat{y}_t = \frac{1}{2}\sum_{t=1}^{T} \hat{y}_t^\top \hat{y}_t \alpha_t \leq \frac{1}{2} D^2 \sum_{t=1}^{T} \alpha_t = \frac{1}{2} D^2 \bar{\alpha}_T = D\sqrt{G_T}.$$

- Right of $B_R(\vec{Q}_T, \vec{g_T})$:

$$\sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t = \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{g_{t,i}^2}{\alpha_{1:t}} = \sum_{t=1}^{T} \frac{D}{2} \frac{\sum_{i=1}^{n} g_{t,i}^2}{\sqrt{G_t}} \leq D\sqrt{G_T}.$$

Proof of Corollary 8 (cont'd)

- In order to make a competitive guarantee, prove a lower bound on the post-hoc optimal bound function B_R. When $\mathcal{F} = \{x | \|x\|_2 \leq D/2\}$, the best post-hoc bound is

$$\min_{\alpha \geq 0} \left(\frac{1}{2} \alpha D^2 + \frac{1}{\alpha} G_T \right) = D\sqrt{2G_T},$$

so conlude the adaptive algorithm is $\sqrt{2}$-competitive for the bound optimization problem.

3.2. Adaptive diagonal regularization

- Define the projection operator,

$$P_{\mathcal{F},A}(u) = \underset{x \in \mathcal{F}}{\operatorname{argmin}} \|A(x-u)\|.$$

Then FTPRL update has an equivalent form as following:

$$x_{t+1} = \underset{x \in \mathcal{F}}{\operatorname{argmin}} (r_{1:t}(x) + g_{1:t}x) \qquad \text{(Original FTPRL)}$$

$$\Leftrightarrow \begin{cases} u_{t+1} & = \operatorname{argmin}_{u \in \mathbb{R}^n} (r_{1:t}(u) + g_{1:t}u) \\ x_{t+1} & = P_{\mathcal{F},A_t}(u_{t+1}) \end{cases} \qquad \text{(Unconstrained optimization)}$$

3.2. Adaptive diagonal regularization

- To derive a algorithm, first construct a closed-form solution to the unconstrained problem.
- Since $r_t(u) = \frac{1}{2}(u - x_t)^\top Q_t(u - x_t)$, we have

$$\frac{\partial r_{1:t}(u)}{\partial u} = Q_{1:t}u - \sum_{\tau=1}^{t} Q_\tau x_\tau .$$

Because u_{t+1} is the optimum of the uncontrained problem, $\left.\frac{\partial r_{1:t}(u)}{\partial u} + g_{1:t}\right|_{u=u_{t+1}} = 0$, hence,

$$u_{t+1} = Q_{1:t}^{-1}\left(\sum_{\tau=1}^{t} Q_\tau x_\tau - g_{1:t}\right).$$

- In this section, set ith entry on the diagonal of $Q_{1:t}$ as

$$\bar{\lambda}_{t,i} = \frac{2}{D_i}\sqrt{\sum_{\tau=1}^{t} g_{\tau_{t,i}^2}}.$$

3.2. Adaptive diagonal regularization

Algorithm 1 FTPRL-Diag

Input: feasible set $\mathcal{F} \subseteq \times_{i=1}^{n}[a_i, b_i]$
Initialize $x_1 = 0 \in \mathcal{F}$
$(\forall i),\ G_i = 0, q_i = 0, \lambda_{0,i} = 0, D_i = b_i - a_i$
for $t = 1$ **to** T **do**
  Play the point x_t, incur loss $f_t(x_t)$
  Let $g_t \in \partial f_t(x_t)$
  for $i = 1$ **to** n **do**
    $G_i = G_i + g_{t,i}^2$
    $\lambda_{t,i} = \frac{2}{D_i}\sqrt{G_i} - \lambda_{1:t-1,i}$
    $q_i = q_i + x_{t,i}\lambda_{t,i}$
    $u_{t+1,i} = (g_{1:t,i} - q_i)/\lambda_{1:t,i}$
  end for
  $A_t = \mathrm{diag}(\sqrt{\lambda_{1:t,1}}, \ldots, \sqrt{\lambda_{1:t,n}})$
  $x_{t+1} = \mathrm{Project}_{\mathcal{F},A_t}(u_{t+1})$
end for

3.2. Adaptive diagonal regularization

Corollary 9 Let $\mathcal{F}$ be a convex feasible set of width D_i in coordinate i. Then, if we run FTPRL with diagonal matrices s.t.

$$(Q_{1:t})_{ii} = \bar{\lambda}_{t,i} = \frac{2}{D_i}\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2},$$

then

$$\text{Regret} \leq 2\sum_{i=1}^{n} D_i\sqrt{\sum_{t=1}^{T} g_{t,i}^2}.$$

- When $\mathcal{F}$ is a hyperrectangle, then this algorithm is $\sqrt{2}$-competitive with the post-hoc optimal choice of Q_t from the $\mathcal{Q}_{\text{diag}} := \{\text{diag}(\lambda_1, \ldots, \lambda_n) | \lambda_i \geq 0\}$.

3.2. Adaptive diagonal regularization

Example: Practical importance of adaptive regularization

- Suppose $\mathcal{F} = \left[-\frac{1}{2}, \frac{1}{2}\right]^n$, then the diameter of $\mathcal{F}$ is $\sqrt{n}$. On each round t, $g_{t,i}$ is 1 w.p. $i^{-\alpha}$ and is 0 o.w., for some $\alpha \in [1, 2)$.
- Then expected regret bound are
  - GD with a global learning rate: $O(\sqrt{nT})$
  - FTPRL-Diag (using Cor. 9 with $D_i = 1$ and Jensen's ineq.):

$$\mathbb{E}\left[\sum_{i=1}^{n} \sqrt{\sum_{t=1} g_{t,i}^2}\right] \leq \sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} \mathbb{E}[g_{t,i}^2]} = \sum_{i=1}^{n} \sqrt{Ti^{-\alpha}} = O(\sqrt{T} \cdot n^{1-\frac{\alpha}{2}})$$

Theorem 10 Let $\mathcal{F}$ be an aribrary feasible set, bounded by a hyperrectangle H^{out} of width W_i in coordinate i; let H^{in} be an hyperrectangle contained by $\mathcal{F}$ of width $w_i > 0$ in coordinate i, i.e.,

$$H^{\text{in}} \subseteq \mathcal{F} \subseteq H^{\text{out}}.$$

Let $\beta = \max_i \frac{W_i}{w_i}$. Then, the FTPRL-Diag is $\sqrt{2}\beta$-competitve with $\mathcal{Q}_{\text{diag}}$ on $\mathcal{F}$.

3.3. A post-hoc bound for diagonal regularization on L_p balls

- Suppose the feasible set $\mathcal{F}$ is an unit L_p ball: $\mathcal{F} = \{x | \|x\|_p \leq 1\}$
- Consider the post-hoc bound optimization problem with $\mathcal{Q} = \mathcal{Q}_{\text{diag}}$.

Theorem 11 For $p > 2$, the optimal regularization matrix for B_R in $\mathcal{Q}_{\text{diag}}$ is not coordiante-constant, except in the degenerate case where $G_i = \sum_{t=1}^{T} g_{t,i}^2$ is the same for all i. However for $p \leq 2$, the optimal regularization matrix in $\mathcal{Q}_{\text{diag}}$ always belongs to $\mathcal{Q}_{\text{const}}$.

3.4. Full matrix regularization on hyperspheres and hyperellipsoids

- In this section, we develop an algorithm for feasible sets $\mathcal{F} \subseteq \{x | \|Ax\|_p \leq 1\}$, where $p \in [1, 2]$ and $A \in S^n_{++}$.
- **Theorem 13** When $\mathcal{F} = \{x | \|Ax\|_2 \leq 1\}$, this algorithm (FTPRL-Scale), is $\sqrt{2}$-competitive with arbitrary S^n_+. For $\mathcal{F} = \{x | \|Ax\|_p \leq 1\}$ with $p \in [1, 2)$ it is $\sqrt{2}$-competitive with $\mathcal{Q}_{\text{diag}}$.

3.4. Full matrix regularization on hyperspheres and hyperellipsoids

Theorem 12 Fix an arbitrary norm $\|\cdot\|$, and define two online linear optimization problem:

1. $\mathcal{I} = (\mathcal{F}, (g_1, \ldots, g_T))$ where $\mathcal{F} = \{x \mid \|Ax\| \leq 1\}$ with $A \in S^n_{++}$
2. $\hat{\mathcal{I}} = (\hat{\mathcal{F}}, (\hat{g}_1, \ldots, \hat{g}_T))$ where $\hat{\mathcal{F}} = \{\hat{x} \mid \|\hat{x}\| \leq 1\}$ and $\hat{g}_t = A^{-1} g_t$.

Then if we run any algorithm dependent only on subgradients on $\hat{\mathcal{I}}$, and it plays $\hat{x}_1, \ldots, \hat{x}_T$, then by playing the corresponding points $x_t = A^{-1}\hat{x}_t$ on $\mathcal{I}$ we achieve identical loss and regret. Furthermore, the post-hoc optimal bound over arbitrary $Q \in S^n_{++}$ is identical for these two instance.

- Using Thm 12, we can now define the adaptive algorithm FTPRL-Scale.

3.4. Full matrix regularization on hyperspheres and hyperellipsoids

Algorithm 2 FTPRL-Scale

Input: feasible set $\mathcal{F} \subseteq \{x \mid \|Ax\| \leq 1\}$, with $A \in S^n_{++}$

Let $\hat{\mathcal{F}} = \{x \mid \|x\| \leq 1\}$

Initialize $x_1 = 0$, $(\forall i)\ D_i = b_i - a_i$

for $t = 1$ **to** T **do**

  Play the point x_t, incur loss $f_t(x_t)$

  Let $g_t \in \partial f_t(x_t)$

  $\hat{g}_t = (A^{-1})^\top g_t$

  $\bar{\alpha} = \sqrt{\sum_{\tau=1}^{t} \sum_{i=1}^{n} \hat{g}_{\tau,i}^2}$

  $\alpha_t = \bar{\alpha} - \alpha_{1:t-1}$

  $q_t = \alpha_t x_t$

  $\hat{u}_{t+1} = (1/\bar{\alpha})(q_{1:t} - g_{1:t})$

  $A_t = (\bar{\alpha} I)^{\frac{1}{2}}$

  $\hat{x}_{t+1} = \text{Project}_{\hat{\mathcal{F}}, A_t}(\hat{u}_{t+1})$

  $x_{t+1} = A^{-1}\hat{x}$

end for

3.4. Full matrix regularization on hyperspheres and hyperellipsoids

Example: FTPRL-Scale has a better bound.

- Let $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$ and $A = \text{diag}(1/a_1, \ldots, 1/a_n)$ with $a_i > 0$. WLOG, assume $\max_i a_i = 1$. Then diameter$(\mathcal{F}) = 2$.
- We compare the regret bound obtained by directly applying the algorithm of Cor. 8 to that of the FTPRL-Scale algorithm.
- By Cor. 8, recalling $G_i = \sum_{t=1}^{T} g_{t,i}^2$, we have

$$\text{Regret} \leq 4\sqrt{\sum_{i=1}^{n} G_i} \tag{3}$$

- Now consider FTPRL-Scale, which uses the transformation of Thm. 12. Applying Cor. 8 to the transformed problem gives

$$\text{Regret} \leq 4\sqrt{\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{g}_{t,i}^2} = 4\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{t=1}^{T} g_{t,i}^2} = 4\sqrt{\sum_{i=1}^{n} a_i^2 G_i}$$

Theorem 11 For $p > 2$, the optimal regularization matrix for B_R in $\mathcal{Q}_{\text{diag}}$ is not coordiante-constant, except in the degenerate case where $G_i = \sum_{t=1}^{T} g_{t,i}^2$ is the same for all i. However for $p \leq 2$, the optimal regularization matrix in $\mathcal{Q}_{\text{diag}}$ always belongs to $\mathcal{Q}_{\text{const}}$.

- Since $\mathcal{F} = \{x | \|x\|_p \leq 1\}$ is symmetric, the optimal post-hoc choice will be in the form as

$$\inf_{Q \in \mathcal{Q}_{\text{diag}}} \max_{y \in \mathcal{F}} (2y^\top Q y) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t.$$

Letting $Q = \text{diag}(\lambda_1, \ldots, \lambda_n)$, we can re-write above optimization problem as

$$\max_{y: \|y\|_p \leq 1} \left(2 \sum_{i=1}^{n} y_i^2 \lambda_i \right) + \sum_{i=1}^{n} \frac{G_i}{\lambda_i}. \tag{4}$$

- For $p \geq 2$, using the change of variable technique and the Hölder inequality, we have

$$\max_{y: \|y\|_p \leq 1} \left(2 \sum_{i=1}^{n} y_i^2 \lambda_i \right) = \max_{z: \|z\|_{\frac{p}{2}} \leq 1} 2 \sum_{i=1}^{n} z_i \lambda_i = 2\|\lambda\|_q,$$

where $q = \frac{p}{p-2}$ (allowing $q = \infty$ for $p = 2$).

- Thus, for $p \geq 2$, the previous bound simplifies to

$$B(\lambda) = 2\|\lambda\|_q + \sum_{i=1}^{n} \frac{G_i}{\lambda_i} \tag{5}$$

1 First suspose $p > 2$.

- Then

$$\Delta B(\lambda)_i := \frac{\partial B(\lambda)}{\partial \lambda_i} = \frac{2}{q}\left(\sum_{i=1}^{n} \lambda_i^q\right)^{\frac{1}{q}-1} \cdot q\lambda_i^{q-1} - \frac{G_i}{\lambda_i^2} = 2\left(\frac{\lambda_i}{\|\lambda\|_q}\right)^{q-1} - \frac{G_i}{\lambda_i^2}.$$

- If $\lambda_1 = \cdots = \lambda_n$, then we have

$$\left(\frac{\lambda_i}{\|\lambda\|_q}\right)^{q-1} = \left(\frac{\lambda_1}{(n\lambda_1^q)^{\frac{1}{q}}}\right)^{q-1} = \left(\frac{1}{n^{\frac{1}{q}}}\right) = n^{\frac{1}{q}-1}.$$

- Hence ith component of the gradient is $2n^{\frac{1}{q}-1} - \frac{G_i}{\lambda_1^2}$, and so if not all the G_i's are equal, some component of the gradient is non-zero! ($\Rightarrow\!\Leftarrow$)

2 For $p \in [1, 2]$,

- it is easy to show that the sol. to Eq. (4) is

$$B_\infty(\lambda) = 2\|\lambda\|_\infty + \sum_{i=1}^{n} \frac{G_i}{\lambda_i}. \tag{6}$$

- The left-term of $B_\infty(\lambda)$ only depend on the largest λ_i, and on the right hand we would like all λ_i as large as possible, a solution of the form $\lambda_1 = \cdots = \lambda_n$ must be optimal.

Theorem 13 The diagonal-constant algorithm analyzed in Cor. 8 is $\sqrt{2}$-competitive with S^n_+ when $\mathcal{F} = \{x | \|x\|_p \leq 1\}$ for $p = 2$, and $\sqrt{2}$-competitive against $\mathcal{Q}_{\text{diag}}$ when $p \in [1, 2)$. Furthermore, when $\mathcal{F} = \{x | \|Ax\|_p \leq 1\}$ with $A \in S^n_{++}$, the FTPRL-Scale algorithm achieves these same competitive guarantees.

1. The results for $\mathcal{Q}_{\text{diag}}$ with $p \in [1, 2)$ follow from Thm 11, 12 and Cor. 8.
2. Consider $p = 2$, $Q \in S^n_{++}$, $\mathcal{F} = \{x | \|x\|_p \leq 1\}$.
   - Then Eq. (6) is tight, so the post-hoc bound for Q is

     $$2 \max_i (\lambda_i) + \sum_{t=1}^{T} g_t^\top (PD^{-1}P^\top) g_t,$$

     where $Q = PDP^\top$, D is a diagonal matrix of positivie eigenvalues and $PP^\top = I$.
     Let $z_t = P^\top g_t$, so each right-hand term is $\sum_{i=1}^n \frac{z_{t,i}^2}{\lambda_i}$. Hence a solution where $D = \alpha I, \alpha > 0$ is optimal.
   - Then we have

     $$B(\alpha) = 2\alpha + \sum_{t=1}^{T} g_t^\top \left(P \left(\frac{1}{\alpha} I \right) P^\top \right) g_t = 2\alpha + \frac{1}{\alpha} \sum_{t=1}^{T} g_t^\top g_t = 2\alpha + \frac{G_T}{\alpha}$$

   - Setting $\alpha = \sqrt{G_T/2}$ produces a minimal post-hoc bound of $2\sqrt{2G_T}$, and the coordinate-constant algorithm has regret bound $4\sqrt{G_T}$.