Supplementary Material for "Multi-label Multiple Kernel Learning by Stochastic Approximation: Application to Visual Object Recognition"

Serhat S. Bucak*
bucakser@cse.msu.edu

Rong Jin*
rongjin@cse.msu.edu

Anil K. Jain*†
jain@cse.msu.edu

Dept. of Comp. Sci. & Eng.*
Michigan State University
East Lansing, MI 48824,U.S.A.

Dept. of Brain & Cognitive Eng.†
Korea University, Anam-dong,
Seoul, 136-713, Korea

Proposition 1. *Eq. (2) is the dual problem of Eq. (1).*

$$\min_{\mathbf{p}\in\mathcal{P}} \min_{\{f_k\in\mathcal{H}(\mathbf{p})\}_{k=1}^m} \left\{\sum_{k=1}^m H_k = \sum_{k=1}^m \left\{\frac{1}{2}|f_k|^2_{\mathcal{H}(\mathbf{p})} + \sum_{i=1}^n \ell\left(y_i^k f_k(\mathbf{x}_i)\right)\right\}\right\}, \tag{1}$$

where $\ell(z) = \max(0, 1-z)$ *and* $\mathcal{H}(\mathbf{p})$ *is a Reproducing Kernel Hilbert Space endowed with kernel* $\kappa(\mathbf{x},\mathbf{x}';\mathbf{p}) = \sum_{a=1}^s p^a \kappa_a(\mathbf{x},\mathbf{x}')$.

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\alpha}\in\mathcal{Q}_1} \left\{\mathcal{L}(\mathbf{p},\boldsymbol{\alpha}) = \sum_{k=1}^m \left\{[\boldsymbol{\alpha}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}^k\circ\mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\alpha}^k\circ\mathbf{y}^k)\right\}\right\}, \tag{2}$$

where $\mathcal{Q}_1 = \left\{\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^m) : \boldsymbol{\alpha}^k \in [0,C]^n, k=1,\ldots,m\right\}$.

Proof. We first rewrite $\ell(z)$ as

$$\ell(z) = \max_{x\in[0,1]} (x - xz),$$

Using the above expression for $\ell(z)$, the second term of H_k can be rewritten as,

$$\sum_{i=1}^n \max_{\alpha_i^k\in[0,C]} \left(\alpha_i^k - \alpha_i^k y_i^k f_k(\mathbf{x}_i)\right),$$

According to von Newman's lemma, we can swith minimization (over f_k) with maximization (over $\boldsymbol{\alpha}$). By taking the minimization over f_k first, we have

$$f_k(x) = \sum_{i=1}^n y_i^k \alpha_i^k \kappa(x_i, x).$$

Finally the problem becomes

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\alpha}\in[0,C]} \left\{\mathcal{L}(\mathbf{p},\boldsymbol{\alpha}) = \sum_{k=1}^m \left\{[\boldsymbol{\alpha}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}^k\circ\mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\alpha}^k\circ\mathbf{y}^k)\right\}\right\}.$$

□

Proposition 2. *Eq. (4) is the dual problem of (3).*

$$\min_{\mathbf{p}\in\mathcal{P}} \min_{\{f_k\in\mathcal{H}(\mathbf{p})\}_{k=1}^m} \max_{1\le k\le m} H_k, \tag{3}$$

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\beta}\in B} \left\{ \mathcal{L}(\mathbf{p},\boldsymbol{\beta}) = \left\{ \sum_{k=1}^m \left\{ [\boldsymbol{\beta}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}^k \circ \mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\beta}^k \circ \mathbf{y}^k) \right\}^{\frac{1}{2}} \right\}^2 \right\}. \tag{4}$$

where

$$B = \left\{ (\boldsymbol{\beta}^1,\dots,\boldsymbol{\beta}^m) : \boldsymbol{\beta}^k \in \mathbb{R}_+^n, k=1,\dots,m, \boldsymbol{\beta}^k \in [0, C\lambda_k]^n \textit{ s.t. } \sum_{k=1}^m \lambda_k = 1 \right\}.$$

Proof. We start by formulating (3) as,

$$\min_{\mathbf{p}\in\mathcal{P}} \min_{\{f_k\in\mathcal{H}(\mathbf{p})\}_{k=1}^m} \min t \tag{5}$$

$$\text{subject to } H_k \le t,\ for k = 1,\dots,m, \tag{6}$$

with extra variable $t \in \mathbb{R}$. Introducing the multiplier λ_k for $H_k \le t$, and using Proposition 1, the Lagrangian is

$$t+\sum_{k=1}^m \lambda_k \left\{ [\boldsymbol{\alpha}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}^k \circ \mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\alpha}^k \circ \mathbf{y}^k) - t \right\} = (1-\mathbf{1^T}\boldsymbol{\lambda})t+\sum_{k=1}^m \lambda_k \left\{ [\boldsymbol{\alpha}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}^k \circ \mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\alpha}^k \circ \mathbf{y}^k) \right\},$$

where $\boldsymbol{\alpha} \in [0, C]^n$. So, the dual function is

$$g(\mathbf{p},\boldsymbol{\beta},\boldsymbol{\lambda}) = \begin{cases} \sum_{k=1}^m \left\{ [\boldsymbol{\beta}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}^k \circ \mathbf{y}^k)^\top \frac{\mathbf{K}(\mathbf{p})}{\lambda_k}(\boldsymbol{\beta}^k \circ \mathbf{y}^k) - t \right\} & \mathbf{1}^\top\boldsymbol{\lambda} = 1 \\ -\infty & \text{otherwise} \end{cases},$$

where $\boldsymbol{\beta}_k = \boldsymbol{\alpha}^k \lambda_k$. Then the dual problem is

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\beta}\in B} \max_{\boldsymbol{\lambda}\in\Lambda} \left\{ \mathcal{L}(\mathbf{p},\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{k=1}^m \left\{ [\boldsymbol{\beta}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}^k \circ \mathbf{y}^k)^\top \frac{\mathbf{K}(\mathbf{p})}{\lambda_k}(\boldsymbol{\beta}^k \circ \mathbf{y}^k) \right\} \right\},$$

where

$$B = \left\{ (\boldsymbol{\beta}^1,\dots,\boldsymbol{\beta}^m) : \boldsymbol{\beta}^k \in \mathbb{R}_+^n, k=1,\dots,m, \boldsymbol{\beta}^k \in [0, C\lambda_k]^n \right\}.$$

Let $\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\beta}\in B} (\boldsymbol{\beta}^k \circ \mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\beta}^k \circ \mathbf{y}^k) = \rho^k$. To eliminate $\boldsymbol{\lambda}$, we rewrite the dual problem as maximization over $\boldsymbol{\lambda}$ for optimal ρ^k. Then, the Lagrangian becomes

$$\max_{\boldsymbol{\lambda}\in\Lambda} -\frac{1}{2}\sum_{k=1}^m \frac{\rho^k}{\lambda_k} + \upsilon(\sum_{k=1}^m \lambda_k - 1).$$

Maximizing over $\boldsymbol{\lambda}$, we get

$$\upsilon = \frac{1}{2}\left\{ \sum_{k=1}^m \sqrt{\rho}_k \right\}^2$$

$$\lambda_k = \frac{\sqrt{\rho_k}}{\sum_{j=1}^m \sqrt{\rho_j}}$$

By eliminating $\boldsymbol{\lambda}$, we obtain the following dual of (3):

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\boldsymbol{\beta}\in B} \left\{ \mathcal{L}(\mathbf{p},\boldsymbol{\beta}) = \left\{ \sum_{k=1}^m \left\{ [\boldsymbol{\beta}^k]^\top \mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}^k \circ \mathbf{y}^k)^\top \mathbf{K}(\mathbf{p})(\boldsymbol{\beta}^k \circ \mathbf{y}^k) \right\}^{\frac{1}{2}} \right\}^2 \right\}.$$

□

Proposition 3. *We define potential functions* $\Phi_p = \frac{\eta_p}{\eta_\gamma} \sum_{a=1}^{s} p^a \ln p^a$ *for* $\mathbf{p}$ *and* $\Phi_\gamma = \sum_{i=1}^{m} \gamma^i \ln \gamma^i$ *for* $\boldsymbol{\gamma}$*, and have the following equations for updating* $\mathbf{p}_t$ *and* $\boldsymbol{\gamma}_t$ *as*

$$p_{t+1}^a = \frac{p_t^a}{Z_t^p} \exp(-\eta_p \nabla_{p^a} \mathcal{L}(\mathbf{p}_t, \gamma_t)), \ \gamma_{t+1}^k = \frac{\gamma_t^k}{Z_t^\gamma} \exp(-\eta_\gamma \nabla_{\gamma^k} \mathcal{L}(\mathbf{p}_t, \gamma_t)), \tag{7}$$

where Z_t^p *and* Z_t^γ *are normalization factors that ensure* $\mathbf{p}_t^\top \mathbf{1} = \boldsymbol{\gamma}_t^\top \mathbf{1} = 1$.

Proof. We denote by $D_{\Phi_{\mathbf{p}}}(\mathbf{p}, \mathbf{p}') : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}_+$ and $D_{\Phi_\gamma}(\gamma, \gamma') : \Gamma \times \Gamma \mapsto \mathbb{R}_+$ the Bregman distance functions for $\mathbf{p}$ and $\boldsymbol{\gamma}$ that are induced by $\Phi_{\mathbf{p}}$ and Φ_γ, respectively. Note that the Bregman distance between z and z' induced by the strictly convex function Φ, denoted by $D_\Phi(z, z')$, is defined as

$$D_\Phi(z, z') = \Phi(z) - \Phi(z') - \nabla\Phi(z')^\top (z - z')$$

Using the Bregman distance function, we introduce two projection operators: $A_{\mathbf{p}}(\mathbf{g}_{\mathbf{p}}; \mathcal{P})$ that projects solution $\mathbf{p}$ into domain $\mathcal{P}$ along the direction $\mathbf{g}_{\mathbf{p}} \in \mathbb{R}^s$ and $B_\gamma(\mathbf{g}_\gamma; \Gamma)$ that projects solution $\boldsymbol{\gamma}$ into domain Γ along the direction $\mathbf{g}_\gamma \in \mathbb{R}^m$. These two operators are defined as follows:

$$A_{\mathbf{p}}(\mathbf{g}_{\mathbf{p}}) = \min_{\mathbf{p}' \in \mathcal{P}} \mathbf{g}_{\mathbf{p}}^\top \mathbf{p}' + D_{\Phi_p}(\mathbf{p}', \mathbf{p}), \ B_{\boldsymbol{\gamma}}(\mathbf{g}_\gamma) = \min_{\gamma' \in \Gamma} \mathbf{g}_\gamma^\top \boldsymbol{\gamma}' + D_{\Phi_\gamma}(\boldsymbol{\gamma}', \boldsymbol{\gamma})$$

Based on the mirror prox method, we can solve the optimization problem in Eq. (5) on the paper iteratively. Given the solution $\mathbf{p}_t$ and $\boldsymbol{\gamma}_t$ of the current iteration, the new solution, denoted by $\mathbf{p}_{t+1}$ and $\boldsymbol{\gamma}_{t+1}$, is computed as

$$p_{t+1} = A_{\mathbf{p}_t}\left(\eta_p \nabla_{\mathbf{p}} \mathcal{L}(\mathbf{p}_t, \boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\right), \quad \boldsymbol{\gamma}_{t+1} = C_{\gamma_t}\left(-\eta_\gamma \nabla_\gamma \mathcal{L}(\mathbf{p}_t, \boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\right), \tag{8}$$

where $\eta_p > 0$ and $\eta_\gamma > 0$ are the step sizes. The two gradients are computed as

$$\begin{aligned} g_i(\mathbf{p}) &= \frac{\partial \mathcal{L}(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\alpha})}{\partial p_i} = -\frac{1}{2} \sum_{k=1}^{m} \gamma_k (\boldsymbol{\alpha}^k \circ \mathbf{y}^k)^\top K_i (\boldsymbol{\alpha}^k \circ \mathbf{y}^k), i = 1, \ldots, s && (9) \\ g_k(\boldsymbol{\gamma}) &= \frac{\partial \mathcal{L}(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\alpha})}{\partial \gamma_k} = [\boldsymbol{\alpha}^k]^\top \mathbf{1} - \frac{1}{2} (\boldsymbol{\alpha}^k \circ \mathbf{y}^k)^\top K(\mathbf{p}) (\boldsymbol{\alpha}^k \circ \mathbf{y}^k) i = k, \ldots, m && (10) \\ & && (11) \end{aligned}$$

By choosing the potential functions as

$$\Phi_{\mathbf{p}} = \frac{\eta_p}{\eta_\gamma} \sum_{a=1}^{s} p_a \ln p_a, \quad \Phi_\gamma = \sum_{a=1}^{m} \gamma_a \ln \gamma_a, \tag{12}$$

we have the following updating rules for $\mathbf{p}_{t+1} = (p_{t+1}^1, \ldots, p_{t+1}^s)$ and $\gamma_{t+1} = (\gamma_{t+1}^1, \ldots, \gamma_{t+1}^m)$

$$\begin{aligned} p_{t+1}^i &= \frac{p_t^i}{Z_t^p} \exp\left(-\eta_\gamma g_i(\mathbf{p}_t)\right), i = 1, \ldots, s && (13) \\ \gamma_{t+1}^i &= \frac{\gamma_t^i}{Z_t^\gamma} \exp\left(\eta_\gamma g_i(\gamma_t)\right), i = 1, \ldots, m && (14) \end{aligned}$$

where Z_t^p and Z_t^γ are defined as

$$Z_t^p = \sum_{i=1}^{s} p_t^i \exp\left(-\eta_\gamma g_i(\mathbf{p}_t)\right) \quad Z_t^\gamma = \sum_{i=1}^{m} \gamma_t^i \exp\left(\eta_\gamma g_i(\gamma_t)\right)$$

□

Theorem 1. *After running Algorithm 1 over T iterations, we have the following inequality for the solution $\widehat{\mathbf{p}}$ and $\widehat{\gamma}$ obtained by Algorithm 1*

$$\mathrm{E}\left[\Delta\left(\widehat{\mathbf{p}},\widehat{\gamma}\right)\right] \leq \frac{1}{\eta_\gamma T}\left(\ln m + \ln s\right) + \eta_\gamma\left(d\frac{m^2}{2\delta^2}\lambda_0^2 n^2 C^4 + n^2 C^2\right),$$

where d is a constant term and $\mathrm{E}[\cdot]$ stands for the expectation over the sampled task indices of all iterations.

Proof. Define

$$\widehat{\mathbf{g}}^p(\mathbf{p}_t,\gamma_t) = (\widehat{g}_1^p(\mathbf{p}_t,\gamma_t),\ldots,\widehat{g}_s^p(\mathbf{p}_t,\gamma_t)),\ \widehat{\mathbf{g}}^\gamma(\mathbf{p}_t,\gamma_t) = (\widehat{g}_1^\gamma(\mathbf{p}_t,\gamma_t),\ldots,\widehat{g}_m^\gamma(\mathbf{p}_t,\gamma_t)).$$

Using the result of variation inequality [1], we have the following inequality for any $\mathbf{p}\in\mathcal{P}$ and $\gamma\in\Gamma$

$$\Delta\left(\mathbf{p}_t,\gamma_t\right) \leq (\mathbf{p}_t-\mathbf{p})^\top\nabla_{\mathbf{p}}\mathcal{L}(\mathbf{p}_t,\gamma_t) - (\gamma_t-\gamma)^\top\nabla_\gamma\mathcal{L}(\mathbf{p}_t,\gamma_t). \tag{15}$$

According to Proposition 1, we have

$$\mathrm{E}_t\left[\widehat{g}^p(\mathbf{p}_t,\gamma_t)\right] = \nabla_{\mathbf{p}}\mathcal{L}(\mathbf{p}_t,\gamma_t),\ \mathrm{E}_t\left[\widehat{g}^\gamma(\mathbf{p}_t,\gamma_t)\right] = \nabla_\gamma\mathcal{L}(\mathbf{p}_t,\gamma_t).$$

We therefore can rewrite Eq. (15) as

$$\mathrm{E}_t\left[\Delta\left(\mathbf{p}_t,\gamma_t\right)\right] \leq \mathrm{E}_t\left[(\mathbf{p}_t-\mathbf{p})^\top\widehat{\mathbf{g}}^p(\mathbf{p}_t,\gamma_t) - (\gamma_t-\gamma)^\top\widehat{\mathbf{g}}^\gamma(\mathbf{p}_t,\gamma_t)\right].$$

From [2] (chapter 11), we know that

$$\eta_\gamma(\mathbf{p}_t-\mathbf{p})^\top\widehat{\mathbf{g}}^p(\mathbf{p}_t,\gamma_t) \leq \mathrm{KL}(\mathbf{p}\|\mathbf{p}_t) - \mathrm{KL}(\mathbf{p}\|\mathbf{p}_{t+1}) + \mathrm{KL}(\mathbf{p}_t\|\mathbf{p}_{t+1}),$$

and

$$-\eta_\gamma(\gamma_t-\gamma)^\top\widehat{\mathbf{g}}^\gamma(\mathbf{p}_t,\gamma_t) \leq \mathrm{KL}(\gamma\|\gamma_t) - \mathrm{KL}(\gamma\|\gamma_{t+1}) + \mathrm{KL}(\gamma_t\|\gamma_{t+1}).$$

Therefore, we have

$$\eta_\gamma\sum_{t=1}^T\Delta\left(\mathbf{p}_t,\gamma_t\right) \leq \mathrm{KL}(\mathbf{p}\|\mathbf{p}_1) + \mathrm{KL}(\gamma\|\gamma_1) + \sum_{t=1}^T\left\{\mathrm{KL}(\mathbf{p}_t\|\mathbf{p}_{t+1}) + \mathrm{KL}(\gamma_t\|\gamma_{t+1})\right\}.$$

We are going to bound each of the three terms on the right hand side of the inequality. First, it is obvious that $\mathrm{KL}(\mathbf{p}\|\mathbf{p}_1) \leq \ln s$ and $\mathrm{KL}(\gamma\|\gamma_1) \leq \ln m$ given both γ_1 and $\mathbf{p}_1$ are uniform distributions. Second, we bound $\mathrm{KL}(\mathbf{p}_t\|\mathbf{p}_{t+1})$ as follows

$$\begin{aligned}
\mathrm{KL}(\mathbf{p}_t|\mathbf{p}_{t+1}) &= \frac{\eta_p}{\eta_\gamma}\left\{\sum_{i=1}^s p_t^i\ln\left(\frac{p_t^i}{p_{t+1}^i}\right)\right\} = \frac{\eta_p}{\eta_\gamma}\left\{\sum_{i=1}^s p_t^i\ln\left(Z_t^p\exp\{\eta_\gamma\widehat{g}_i^p\}\right)\right\} \\
&= \frac{\eta_p}{\eta_\gamma}\left\{\sum_{i=1}^s p_t^i\eta_\gamma\widehat{g}_i^p(\mathbf{p}_t,\gamma_t) + \sum_{i=1}^s p_t^i\ln(Z_t^p)\right\} \\
&= \frac{\eta_p}{\eta_\gamma}\left\{\sum_{i=1}^s p_t^i\eta_\gamma\widehat{g}_i^p(\mathbf{p}_t,\gamma_t) + \sum_{i=1}^s p_t^i\ln\left(\sum_{j=1}^s p_t^j\exp\left[-\eta_\gamma\widehat{g}_j^p(\mathbf{p}_t,\gamma_t)\right]\right)\right\} \\
&= \frac{\eta_p}{\eta_\gamma}\left\{-\left(-\eta_\gamma E\left[\widehat{g}_i^p\right]\right) + \ln\left(E\left[\exp\left(-\eta_\gamma\widehat{g}_j^p(\mathbf{p}_t,\gamma_t)\right)\right]\right)\right\} \\
&\leq \frac{\eta_p}{\eta_\gamma}\left\{\frac{\eta_\gamma^2}{2}\max_{1\leq i\leq s}\left[\widehat{g}_i^p(\mathbf{p}_t,\gamma_t)\right]^2\right\} = \frac{c\eta_\gamma^2}{2}|\widehat{\mathbf{g}}^p(\mathbf{p}_t,\gamma_t)|_\infty^2,
\end{aligned}$$

where the inequality follows directly from the Hoeffiding inequality, and c is a constant such that $\eta_p = c\eta_\gamma$. Similarly, we have $\mathrm{KL}(\gamma_t\|\gamma_{t+1}) \leq \frac{\eta_\gamma^2}{2}|\widehat{\gamma}^\gamma(\mathbf{p}_t,\gamma_t)|_\infty^2$.

By combining the above results together, we have

$$\eta_\gamma \mathrm{E}\left[\sum_{t=1}^{T} \Delta\left(\mathbf{p}_t, \gamma_t\right)\right] \leq \ln m + \ln s + \eta_\gamma^2 \sum_{t=1}^{T} \mathrm{E}\left[c|\widehat{\mathbf{g}}^p(\mathbf{p}_t, \gamma_t)|_\infty^2 + |\widehat{\mathbf{g}}^\gamma(\mathbf{p}_t, \gamma_t)|_\infty^2\right]$$

Using Eq. (8) on the paper, we can bound $|\widehat{\mathbf{g}}^p(\mathbf{p}_t, \gamma_t)|_\infty$ as follows

$$\begin{aligned}
|\widehat{\mathbf{g}}^p(\mathbf{p}_t, \gamma_t)|_\infty &= \max_{1\leq a\leq s} \widehat{g}_a^p(\mathbf{p}_t, \gamma_t) \\
&= \max_{1\leq a\leq s} \left| -\frac{1}{2}(\boldsymbol{\alpha}^{j_t} \circ \mathbf{y}^{j_t})^\top \mathbf{K}^a (\boldsymbol{\alpha}^{j_t} \circ \mathbf{y}^{j_t}) \right| \\
&\leq \frac{1}{2}(C\mathbf{1})^\top \mathbf{V}\mathbf{D}\mathbf{V}^{-1}(C\mathbf{1})] \leq \frac{\lambda_0}{2}(C\mathbf{1})^\top \mathbf{V}\mathbf{I}\mathbf{V}^{-1}(C\mathbf{1}) = \frac{\lambda_0}{2}(C\mathbf{1})^\top \mathbf{I}(C\mathbf{1}) \\
&\leq \frac{1}{2}nC^2\lambda_0,
\end{aligned}$$

where $\mathbf{K} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ is the eigendecomposition of the PSD matrix $\mathbf{K}$, $\lambda_0 = \max\limits_{1\leq a\leq s} \lambda_{\max}(\mathbf{K}^a)$, and $\lambda_{\max}(\mathbf{Z})$ stands for the maximum eigenvalue of matrix $\mathbf{Z}$. Similarly, by using Eq. (9) on the paper we can bound $|\widehat{\mathbf{g}}^\gamma(\mathbf{p}_t, \gamma_t)|_\infty$ as

$$|\widehat{\mathbf{g}}^\gamma(\mathbf{p}_t, \gamma_t)|_\infty = \max_{1\leq k\leq m} \widehat{g}_k^\gamma(\mathbf{p}_t, \gamma_t) \leq \frac{m}{\delta} \max\left(nC, \frac{\lambda_0}{2}nC^2\right).$$

Next, we have the bound simplified as

$$\mathrm{E}\left[\sum_{t=1}^{T} \Delta\left(\mathbf{p}_t, \gamma_t\right)\right] \leq \frac{1}{\eta_\gamma}\left(\ln m + \ln s\right) + \eta_\gamma T\left(d\frac{m^2}{2\delta^2}\lambda_0^2 n^2 C^4 + n^2 C^2\right),$$

where d is a constant. We complete the proof by using the fact $\Delta(\mathbf{p}, \gamma)$ is jointly convex in both $\mathbf{p}$ and γ and therefore $\sum_{t=1}^{T} \Delta\left(\mathbf{p}_t, \gamma_t\right) \geq T\Delta\left(\widehat{\mathbf{p}}, \widehat{\gamma}\right)$. □

Corollary 1. *With $\delta = m^{\frac{2}{3}}$ and $\eta_\gamma = \frac{1}{n}m^{-\frac{1}{3}}\sqrt{(\ln m)/T}$, after running Algorithm 1 (on the paper) over T iterations, we have $\mathrm{E}[\Delta(\widehat{\mathbf{p}}, \widehat{\gamma})] \leq O(m^{1/3}\sqrt{(\ln m)/T})$ in terms of m and T.*

References

[1] A. Nemirovski, “Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems,” *SIAM Journal on Optimization*, vol. 15, pp. 229–251, 2004.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games.* Cambridge University Press, 2006.