Online Appendix 1
"Flexible High-Dimensional Classification Machines and Their Asymptotic Properties"

Xingye Qiao QIAO@MATH.BINGHAMTON.EDU
Department of Mathematical Sciences
Binghamton University
State University of New York
Binghamton, NY 13902-6000, USA

Lingsong Zhang LINGSONG@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Editor: Massimiliano Pontil

This document provides some additional details for the main paper, "Flexible High-dimensional Classification Machines and Their Asymptotic Properties." It includes how we implement the FLAME machine with pre-defined θ, detailed proofs for several theorems and propositions, and figures for additional simulations.

1. Implementation

In order to implement the FLAME algorithm, we introduce several new notations. Let S_{d+1} be a second order cone in the $d+1$ dimensional space, $S_{d+1} = \left\{(t_0, t_1, \cdots, t_d)' : t_0 \geq \sqrt{\sum_{i=1}^{d} t_i^2}\right\}$. Note that r_i and $1/r_i$ can be substituted by three axillary variables ρ_i, σ_i and τ_i which satisfy $\rho_i + \sigma_i = r_i$, $\rho_i - \sigma_i = 1/r_i$, and $\tau_i = 1$. Then $\rho_i^2 = \sigma_i^2 + \tau_i^2$, and thus $(\rho_i, \sigma_i, \tau_i) \in S_3$. Let $w = 1$, then $(w; \boldsymbol{\omega}) \in S_{d+1}$ since $\|\boldsymbol{\omega}\| \leq 1$. Let $\eta_i \geq 0$, $\varphi_i \geq 0$, where φ_i and η_i can be viewed as the positive and negative parts of $\left(\frac{1}{r_i} + C\xi_i - \theta\sqrt{C}\right)$, *i.e.*, $\varphi_i - \eta_i = \left(\frac{1}{r_i} + C\xi_i - \theta\sqrt{C}\right)$. With the reparameterization above, FLAME can be viewed as the following optimization problem:

$$\begin{aligned} \min_{\beta, w, \boldsymbol{\omega}, \rho_i, \sigma_i, \tau_i, \xi_i, \eta_i, \varphi_i} & \sum_{i=1}^{n} \varphi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{x}_i^T \boldsymbol{\omega} + \beta) + \xi_i - \rho_i - \sigma_i = 0 \\ & \rho_i - \sigma_i + C\xi_i - \theta\sqrt{C} + \eta_i - \varphi_i = 0 \\ & w = 1 \\ & \tau_i = 1 \\ \text{and } (w; \boldsymbol{\omega}) \in S_{d+1}, & (\rho_i, \sigma_i, \tau_i)' \in S_3, \xi_i \geq 0, \eta_i \geq 0, \varphi_i \geq 0. \end{aligned}$$

Therefore, all the constraints can be converted to linear forms, all the variables are either nonnegative, free, or in second order cones, and the objective function is linear. Such problem is called Second Order Cone Programming (SOCP), and can be efficiently solved by softwares such as SDPT3 (Toh et al., 1999; Tütüncü et al., 2003).

2. Proof to Theorem 1

It suffices to show that $s(\boldsymbol{\omega}_k, \beta_k, \theta_k) \geq s(\boldsymbol{\omega}_{k+1}, \beta_{k+1}, \theta_{k+1})$. First, $s(\boldsymbol{\omega}_k, \beta_k, \theta_k) \geq s(\boldsymbol{\omega}_k, \beta_k, \theta_{k+1})$ due to the definition of θ_{k+1} and that $\theta_k \leq \theta_{k+1}$. Then $s(\boldsymbol{\omega}_k, \beta_k, \theta_{k+1}) \geq s(\boldsymbol{\omega}_{k+1}, \beta_{k+1}, \theta_{k+1})$ since $\boldsymbol{\omega}_{k+1}$ and β_{k+1} minimize $s(\boldsymbol{\omega}, \beta, \theta_{k+1})$. ■

3. Proof to Proposition 2

For any $\boldsymbol{x}$, denote $p(\boldsymbol{x}) = \mathrm{P}(Y = +1 \mid \boldsymbol{X} = \boldsymbol{x})$. The conditional risk is $R(f) \equiv \mathbb{E}[L(Yf(\boldsymbol{X}), \theta) \mid \boldsymbol{X} = \boldsymbol{x}] = L(f(\boldsymbol{x}), \theta)p(\boldsymbol{x}) + L(-f(\boldsymbol{x}), \theta)(1 - p(\boldsymbol{x}))$. For simplicity, we write $L(f(\boldsymbol{x}), \theta)$ as $L(f)$. Thereby, $R(f) = L(f)p(\boldsymbol{x}) + L(-f)(1 - p(\boldsymbol{x}))$.

We can see that for fixed $p(\boldsymbol{x}) \in (0, 1)$, $R(f)$ is continuous and convex, and differentiable everywhere except the points $\pm 1/(\theta\sqrt{C})$. Thus we find the critical point $f^{**}(\boldsymbol{x})$ by solving $\partial_- R(f) = 0$, where $\partial_- R(f) = \partial_- L(f)p(\boldsymbol{x}) + \partial_-[L(-f)](1 - p(\boldsymbol{x}))$. The FLAME loss is

$$L(f) = \begin{cases} (2-\theta)\sqrt{C} - Cf & \text{if } f \leq \frac{1}{\sqrt{C}} \\ \frac{1}{f} - \theta\sqrt{C} & \text{if } \frac{1}{\sqrt{C}} < f \leq \frac{1}{\theta\sqrt{C}} \\ 0 & \text{otherwise.} \end{cases}$$

So direct calculation gives us that

$$\partial_- L(f) = \begin{cases} -C & \text{if } f \leq \frac{1}{\sqrt{C}} \\ -\frac{1}{f^2} & \text{if } \frac{1}{\sqrt{C}} < f \leq \frac{1}{\theta\sqrt{C}} \\ 0 & \text{otherwise,} \end{cases} \text{ and } \partial_-[L(-f)] = \begin{cases} C & \text{if } f \geq -\frac{1}{\sqrt{C}} \\ \frac{1}{f^2} & \text{if } -\frac{1}{\theta\sqrt{C}} < f \leq -\frac{1}{\sqrt{C}} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, if $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} \neq 1$, the solution to $\partial_- R(f) = 0$ has to be either $-\frac{1}{\theta\sqrt{C}} < f \leq -\frac{1}{\sqrt{C}}$ or $\frac{1}{\sqrt{C}} < f \leq \frac{1}{\theta\sqrt{C}}$. In the former case, $f^{**} = -\frac{1}{\sqrt{C}}\sqrt{\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})}}$ if $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} > 1$. In the latter case, $f^{**} = +\frac{1}{\sqrt{C}}\sqrt{\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}}$ if $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} < 1$. If $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} = 1$, $\partial_- R(f^{**}) = 0$ for any $f^{**} \in [-\frac{1}{\sqrt{C}}, \frac{1}{\sqrt{C}}]$.

It is easy to check that when the solution to $\partial_- R(f) = 0$ exists, $R(f^{**})$ is less than both $R(1/(\theta\sqrt{C}))$ and $R(-1/(\theta\sqrt{C}))$, and f^{**} is the minimizer of $R(f)$.

Finally, when the solution to $\partial_- R(f) = 0$ does not exist, that is, when $\sqrt{\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})}} > \frac{1}{\theta}$ or $\sqrt{\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}} > \frac{1}{\theta}$, we can see that $-\frac{1}{\theta\sqrt{C}}$ is the minimizer when $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} > 1$ and $\frac{1}{\theta\sqrt{C}}$ is the minimizer when $\frac{1-p(\boldsymbol{x})}{p(\boldsymbol{x})} < 1$.

Therefore, the minimizer f^* satisfies $\mathrm{sign}(f^*) = \mathrm{sign}(p(\boldsymbol{x}) - 0.5) = \mathrm{sign}(2p(\boldsymbol{x}) - 1) = \mathrm{sign}(p(\boldsymbol{x}) - (1 - p(\boldsymbol{x}))) = \mathrm{sign}(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} - 1)$. ■

4. Proof to Proposition 3

From the proof to Proposition 2 above, we have that f^*, the minimizer of $R(f) = L(f)\eta + L(-f)(1-\eta)$ with $\eta = p(\boldsymbol{x})$ is

$$f^*(\eta) = \begin{cases} -\frac{1}{\theta\sqrt{C}}, & \text{if } \sqrt{\frac{1-\eta}{\eta}} \geq \frac{1}{\theta},\ i.e.,\ \eta \leq \frac{\theta^2}{1+\theta^2}, \\ -\frac{1}{\sqrt{C}}\sqrt{\frac{1-\eta}{\eta}}, & \text{if } 1 < \sqrt{\frac{1-\eta}{\eta}} < \frac{1}{\theta},\ i.e.,\ \frac{\theta^2}{1+\theta^2} < \eta < 1/2, \\ 0, & \text{if } \eta = 1/2, \\ \frac{1}{\sqrt{C}}\sqrt{\frac{\eta}{1-\eta}}, & \text{if } 1 < \sqrt{\frac{\eta}{1-\eta}} < \frac{1}{\theta},\ i.e.,\ 1/2 < \eta < \frac{1}{1+\theta^2}, \\ \frac{1}{\theta\sqrt{C}}, & \text{if } \sqrt{\frac{\eta}{1-\eta}} \geq \frac{1}{\theta},\ i.e.,\ \eta \geq \frac{1}{1+\theta^2}. \end{cases}$$

Direct calculations lead to that

$$H(\eta) = R(f^*) = \begin{cases} \sqrt{C}\eta(2 + \frac{1}{\theta} - \theta), & \text{if } \eta \leq \frac{\theta^2}{1+\theta^2}, \\ \sqrt{C}[2\eta - \theta + 2\sqrt{\eta(1-\eta)}], & \text{if } \frac{\theta^2}{1+\theta^2} < \eta < 1/2, \\ \sqrt{C}(2-\theta), & \text{if } \eta = 1/2, \\ \sqrt{C}[2(1-\eta) - \theta + 2\sqrt{\eta(1-\eta)}], & \text{if } 1/2 < \eta < \frac{1}{1+\theta^2}, \\ \sqrt{C}(1-\eta)(2 + \frac{1}{\theta} - \theta), & \text{if } \eta \geq \frac{1}{1+\theta^2}. \end{cases}$$

The assertion follows Proposition 2, and Theorem 1 and Theorem 2 of Bartlett et al. (2006). ■

5. Proof to Theorem 4

Since there are infinitely many negative class samples, it is reasonable to assume that the classification boundary is pushed closer to the minority positive class, and therefore, the functional margin $u_i = y_i f(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for the ith vector from the minority positive class is small and its DWD loss is $2\sqrt{C} - Cu_i = 2\sqrt{C} - Cf(\boldsymbol{x}_i)$. Similarly, the DWD loss for the jth vector from the majority negative class is $1/[y_j f(\boldsymbol{x}_j)] = -1/f(\boldsymbol{x}_j)$. The objective function for DWD is therefore equivalent to $\frac{1}{n_+ + n_-}\left\{\sum_{i=1}^{n_+}[2\sqrt{C} - C(\boldsymbol{x}_i^T\boldsymbol{\omega} + \beta)] - \sum_{j=1}^{n_-}\frac{1}{\boldsymbol{x}_j^T\boldsymbol{\omega} + \beta}\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2$.

The second term inside the curly bracket above can be approximated by $n_- \int \frac{1}{\boldsymbol{x}^T\boldsymbol{\omega} + \beta} dF_-(\boldsymbol{x})$ where $F_-(\cdot)$ is the conditional cumulative distribution function for the negative class. The objective function is therefore

$$l_D = \frac{1}{n_+ + n_-}\left\{\sum_{i=1}^{n_+}[2\sqrt{C} - C(\boldsymbol{x}_i^T\boldsymbol{\omega} + \beta)] - n_- \int \frac{1}{\boldsymbol{x}^T\boldsymbol{\omega} + \beta} dF_-(\boldsymbol{x})\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2$$

Before we continue, we need the definition that a distribution has a point surrounded (Owen, 2007).

Definition 1 *The distribution F on $\mathbb{R}^d$ has the point $\boldsymbol{x}_*$ surrounded if*

$$\int_{(\boldsymbol{x}-\boldsymbol{x}_*)'\boldsymbol{\omega} > \epsilon} dF(\boldsymbol{x}) > \delta,$$

for some $\delta > 0$, some $\epsilon > 0$ and all $\boldsymbol{\omega} \in \mathbb{R}^d$ with $\|\boldsymbol{\omega}\| = 1$. Consequentially, if F has $\boldsymbol{x}_*$ surrounded, then there exist γ satisfying

$$\inf_{\|\boldsymbol{\omega}\|=1} \int_{(\boldsymbol{x}-\boldsymbol{x}_*)'\boldsymbol{\omega}>0} dF(\boldsymbol{x}) > \gamma \geq 0. \tag{1}$$

We observe that

$$\begin{aligned}
\frac{\partial l_D}{\partial \beta} &= \frac{1}{n_+ + n_-}[-n_+ C + n_- \int (\boldsymbol{x}^T\boldsymbol{\omega} + \beta)^{-2} dF_-(\boldsymbol{x})] \\
&\geq \frac{1}{n_+ + n_-}[-n_+ C + n_- \int_{(\boldsymbol{x}-\overline{\boldsymbol{x}}_+)'\boldsymbol{\omega}\geq 0} ((\boldsymbol{x}-\overline{\boldsymbol{x}}_+)^T\boldsymbol{\omega} + \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega} + \beta)^{-2} dF_-(\boldsymbol{x})] \\
&\geq \frac{1}{n_+ + n_-}[-n_+ C + n_- \int_{\boldsymbol{x}^T\boldsymbol{\omega}\geq 0} (\overline{\boldsymbol{x}}_+^T\boldsymbol{\omega} + \beta)^{-2} dF_-(\boldsymbol{x})] \\
&\geq \frac{1}{n_+ + n_-}[-n_+ C + n_- \frac{\gamma}{(\overline{\boldsymbol{x}}_+^T\boldsymbol{\omega} + \beta)^2}]
\end{aligned}$$

Now suppose that $-\sqrt{\frac{n_-\gamma}{n_+ C}} < \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega} + \beta < 0$, then $\frac{n_-\gamma}{(\overline{\boldsymbol{x}}_+^T\boldsymbol{\omega}+\beta)^2} > n_+ C$ and $\partial l_D/\partial\beta > 0$. Given the fact that l_D is a strictly convex function, the minimizer $\widehat{\beta} < -\sqrt{\frac{n_-\gamma}{n_+ C}} - \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega}$. ■

6. Proof to Theorem 5

Again, with the imbalance assumption we assume that the functional margins for the minority positive class are always greater than 0. Note that the penalized empirical loss for the FLAME machine is approximated by

$$\begin{aligned}
l_F =& \frac{1}{n_+ + n_-}\left\{\sum_i^{n_+}\left[2\sqrt{C} - C(\boldsymbol{x}_i^T\boldsymbol{\omega} + \beta) - \theta\sqrt{C}\right] + \right. \\
& \left. n_- \int \left(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega} + \beta} - \theta\sqrt{C}\right)_+ dF_-(\boldsymbol{x})\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2
\end{aligned}$$

Let $g_j^* = -(\boldsymbol{x}_j^T\boldsymbol{\omega}^* + \beta^*)$, $j = 1, 2, \cdots n_-$ be the functional margins for the negative class. Because $\frac{1}{g^*_{(n_+)}\sqrt{C}} = \theta^*$, that is, the reduced loss for the jth sample is greater than or equal to 0, $\frac{1}{g^*_{(n_+)}} - \theta^*\sqrt{C} = 0$, we observe that $1/g^*_{(n_+)}$ is the n_+-th greatest among all the function margins of the negative class $1/g_j^* = -1/(\boldsymbol{x}_j^T\boldsymbol{\omega}^* + \beta^*)$. Thus there are at most n_+ samples whose reduced losses that are ≥ 0. Assume that there are $n^o \leq n_+$ such samples.

For a random sample $(\boldsymbol{X}, Y)$ from the negative class, let E be the event that $(Y(\boldsymbol{X}^T\boldsymbol{\omega}^* + \beta^*))^{-1} \geq \theta^*\sqrt{C}$. From the argument above, P(E) is approximately n^o/n_-.

Then the integration $\int \left(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C}\right)_+ dF_-(\boldsymbol{x})$ equals

$$\begin{aligned}
&\mathbb{E}\left[(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C})_+ \mid \overline{E}\right] \mathrm{P}(\overline{E}) + \mathbb{E}\left[(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C})_+ \mid E\right] \mathrm{P}(E) \\
\approx &\mathbb{E}\left[0 \mid \overline{E}\right](1-\frac{n^o}{n_-}) + \mathbb{E}\left[(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C}) \mid E\right]\frac{n^o}{n_-} \\
= &\mathbb{E}\left[(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C}) \mid E\right]\frac{n^o}{n_-}
\end{aligned}$$

We then have

$$\begin{aligned}
l_F = &\frac{1}{n_+ + n_-}\left\{\sum_i^{n_+}\left[2\sqrt{C} - C(\boldsymbol{x}_i^T\boldsymbol{\omega}+\beta) - \theta\sqrt{C}\right] + \right. \\
&\left. n^o\int\left(-\frac{1}{\boldsymbol{x}^T\boldsymbol{\omega}+\beta} - \theta\sqrt{C}\right)dF_-(\boldsymbol{x} \mid E)\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2
\end{aligned}$$

Here, $dF_-(\boldsymbol{x} \mid E)$ is the conditional distribution function of $\boldsymbol{X}$ for the negative class given event E.

Setting $\partial l_F/\partial\beta = 0 = \frac{1}{n_+ + n_-}\left\{-Cn_+ + n^o\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}dF_-(\boldsymbol{x} \mid E)\right\}$, we have $\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}dF_-(\boldsymbol{x} \mid E) = C\frac{n_+}{n^o}$.

Setting $\partial l_F/\partial\boldsymbol{\omega} = \mathbf{0} = \frac{1}{n_+ + n_-}\left\{-Cn_+\overline{\boldsymbol{x}}_+ + n^o\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}\boldsymbol{x}dF_-(\boldsymbol{x} \mid E)\right\} + \lambda\boldsymbol{\omega}^*$, we have $\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}\boldsymbol{x}dF_-(\boldsymbol{x} \mid E) = -\frac{n_+ + n_-}{n^o}\lambda\boldsymbol{\omega}^* + C\frac{n_+}{n^o}\overline{\boldsymbol{x}}_+$. And furthermore,

$$\frac{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}\boldsymbol{x}dF_-(\boldsymbol{x} \mid E)}{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}dF_-(\boldsymbol{x} \mid E)} = -\frac{n_+ + n_-}{n_+}\frac{\lambda}{C}\boldsymbol{\omega}^* + \overline{\boldsymbol{x}}_+ = -(1+m)\frac{\lambda}{C}\boldsymbol{\omega}^* + \overline{\boldsymbol{x}}_+.$$

That is,

$$\boldsymbol{\omega}^* = \frac{C}{(1+m)\lambda}\left[\overline{\boldsymbol{x}}_+ - \frac{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}\boldsymbol{x}dF_-(\boldsymbol{x} \mid E)}{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}dF_-(\boldsymbol{x} \mid E)}\right]$$

■

7. Proof to Theorem 6

For simplicity we use the original SVM formulation with the Hinge loss function instead of the FLAME formulation. The objective function for SVM is equivalent to

$$\begin{aligned}
l_S = &\frac{1}{n_+ + n_-}\left\{\sum_{i=1}^{n_+}[1 - (\boldsymbol{x}_i^T\boldsymbol{\omega}+\beta)] + n_-\int\left[1 + (\boldsymbol{x}^T\boldsymbol{\omega}+\beta)\right]_+ dF_-(\boldsymbol{x})\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2 \\
= &\frac{1}{n_+ + n_-}\left\{\sum_{i=1}^{n_+}[1 - (\boldsymbol{x}_i^T\boldsymbol{\omega}+\beta)] + \right. \\
&\left. n_-\int\left[1 + (\boldsymbol{x}^T\boldsymbol{\omega}+\beta)\right]\mathbb{1}_{\{1+\boldsymbol{x}^T\boldsymbol{\omega}+\beta>0\}}dF_-(\boldsymbol{x})\right\} + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2 \qquad (2)
\end{aligned}$$

Setting $\partial l_S/\partial\beta = 0$, we have

$$\frac{\partial l_S}{\partial \beta} = \frac{1}{n_+ + n_-}\left\{-n_+ + n_- \mathrm{P}(G; \boldsymbol{\omega}, \beta)\right\} = 0,$$

where the derivative of the second term in the curly bracket in (2) is due to Lemma 2 of Koo et al. (2008) under a mild condition that the conditional class densities are continuous and have finite second moments. This leads to

$$\Rightarrow \mathrm{P}(G; \widehat{\boldsymbol{\omega}}, \widehat{\beta}) = \frac{n_+}{n_-} = \frac{1}{m}.$$

Similarly,

$$\begin{aligned}
\frac{\partial l_S}{\partial \boldsymbol{\omega}} =& \frac{1}{n_+ + n_-}\left\{-\sum_{i=1}^{n_+} \boldsymbol{x}_i + n_- \int \boldsymbol{x}\mathbb{1}_{\{1+\boldsymbol{x}^T\boldsymbol{\omega}+\beta>0\}} dF_-(\boldsymbol{x})\right\} + \lambda\boldsymbol{\omega} \\
=& \frac{1}{n_+ + n_-}\left\{-n_+\overline{\boldsymbol{x}}_+ + n_- \int \boldsymbol{x}\mathbb{1}_{\{1+\boldsymbol{x}^T\boldsymbol{\omega}+\beta>0\}} dF_-(\boldsymbol{x})\right\} + \lambda\boldsymbol{\omega} \\
\Rightarrow \widehat{\boldsymbol{\omega}} =& \frac{1}{(n_+ + n_-)\lambda}\left\{n_+\overline{\boldsymbol{x}}_+ - n_- \int \boldsymbol{x}\mathbb{1}_{\{1+\boldsymbol{x}^T\widehat{\boldsymbol{\omega}}+\widehat{\beta}>0\}} dF_-(\boldsymbol{x})\right\} \\
=& \frac{1}{(n_+ + n_-)\lambda}\left\{n_+\overline{\boldsymbol{x}}_+ - n_- \int \boldsymbol{x} dF_-(\boldsymbol{x}|G)\mathrm{P}(G; \widehat{\boldsymbol{\omega}}, \widehat{\beta})\right\} \\
=& \frac{1}{(n_+ + n_-)\lambda}\left\{n_+\overline{\boldsymbol{x}}_+ - n_- \int \boldsymbol{x} dF_-(\boldsymbol{x}|G)\frac{1}{m}\right\} \\
=& \frac{1}{(n_+ + n_-)\lambda}\left\{n_+\overline{\boldsymbol{x}}_+ - n_+ \int \boldsymbol{x} dF_-(\boldsymbol{x}|G)\right\} \\
=& \frac{1}{(1+m)\lambda}\left\{\overline{\boldsymbol{x}}_+ - \int \boldsymbol{x} dF_-(\boldsymbol{x}|G)\right\}.
\end{aligned}$$

■

8. Classification Boundaries for HDLSS Data

The geometric representation in Hall et al. (2005) leads to some theoretical properties of several binary classifiers. In particular, as $d \to \infty$, the positive class and negative class converge to two $(n_+ - 1)$ and $(n_- - 1)$ simplices with random rotation. Note that the (normalized) pairwise distances between observations within each class are the same, and the (normalized) distances between any two observations from two different classes are the same as well. The geometric representation for SVM and DWD in Hall et al. (2005) is summarized as the follows.

1. **SVM**: It was shown that the linear SVM hyperplane projected to the $(N-1)$-dimensional subspace that is generated by the N data vectors is given asymptotically by the unique $(N-2)$-dimensional hyperplane that bisects each of the edges of length l in the N-polyhedron formed by the N data vectors. There are $n_+ \times n_-$ such edges. Let O^+ be the centroid of the $(n_+ - 1)$-simplex $\mathcal{X}^+(d)$ and O^- the centroid of the $(n_- - 1)$-simplex $\mathcal{X}^-(d)$. It can be further shown that the SVM hyperplane bisects the line segment between O^+ and O^-.

2. **DWD**: The case of DWD is a little different, especially in the case where $n_+ \ll n_-$ (or $m \gg 1$), which is our main focus here. We assume that the DWD hyperplane intersects O^+O^- at point P. It can be shown that the two simplices, the DWD hyperplane, and the SVM hyperplane, are all orthogonal to O^+O^-. Thus all the vertices in the simplex $\mathcal{X}^+$ are equally distanced from the DWD hyperplane. Such distance is denoted by a. Similarly, all the vertices in the simplex $\mathcal{X}^-$ are equally distanced from the DWD hyperplane by b. The general version DWD hyperplane minimizes the sum of the reciprocals of the distances of data vectors to the hyperplane, $(n_+/a + n_-/b)$, with the constraint that $a+b$ equals to a constant (determined by μ, σ, τ, n_+, n_-, and d). A simple calculus practice reveals that $a/b = (n_+/n_-)^{1/2}$.

For the General FLAME case, we need to learn how the hyperplane moves from the point determined by $a/b = (n_+/n_-)^{1/2}$ on O^+O^- to the midpoint of O^+O^- as θ grows from 0 (DWD) to 1 (SVM). First, we consider the general version of FLAME which seeks to minimize the sum of losses for all data points, $\sum \left(1/u - \theta\sqrt{C}\right)_+$, where the functional margin u is either a or b for samples from the positive or the negative classes respectively. When $\theta = 0$, the FLAME hyperplane is determined by $a/b = (n_+/n_-)^{1/2} = m^{-1/2} < 1$. Thus $b > a$, that is, the hyperplane is closer to the minority class. We renamed them as a^0 and b^0 where the superscript "0" represents the value of θ.

When $\theta > 0$ but smaller than $1/(b^0\sqrt{C})$, then the hyperplane does not move. This is because that the new loss for each data vector becomes $1/a^0 - \theta\sqrt{C}$ or $1/b^0 - \theta\sqrt{C}$ since both are greater than 0. The additional term "$-\theta\sqrt{C}$" does not change the minimizer and thus $a^{\theta_1}/b^{\theta_1} = (n_+/n_-)^{1/2}$ does remain unchanged.

If we keep increasing θ so that it becomes greater than $1/(b^0\sqrt{C})$, then if the hyperplane does not move, then the loss for the majority class becomes 0. In this case, there is space for improvement: the hyperplane would move gradually towards the majority class, because this can make the loss on the minority class smaller while keeping the loss on the majority zero. The FLAME hyperplane is determined by $b = 1/(\theta\sqrt{C})$.

Finally, as θ increases, the distance a increases and the distance b decreases, until at a point where $a = b$, and both $1/a - \theta\sqrt{C} = 1/b - \theta\sqrt{C} < 0$. After this point, further increase of θ will not change the position of the FLAME hyperplane which will remain at the midpoint of O^+O^-.

The derivation above assumes the distance between the two simplices are reasonable large, at least greater than $2/\sqrt{C}$. This is not difficult to achieve because we choose C to be a large number.

In summary, the intersection P of the FLAME hyperplane and O^+O^- stays closer to the minority class, and remains still as θ is small. When θ increases, the boundary moves towards the majority class, until reaching the midpoint of O^+O^-. This explains the simulation performance we observed in Figures A.2 and A.3, and Figure 6 in the main paper. We use a toy example and show the position of the FLAME hyperplane moving as θ increases in Figure A.1 in the same fashion we discussed above.

It is worth noting that the value of DWD/FLAME in terms of reducing overfitting is maximal when the dimension is greater than, but close to, the sample size. This is when data-piling starts to appear in SVM but not yet in DWD. Marron et al. (2007) showed some videos about such phenomenon. As a matter of fact, according to the geometric

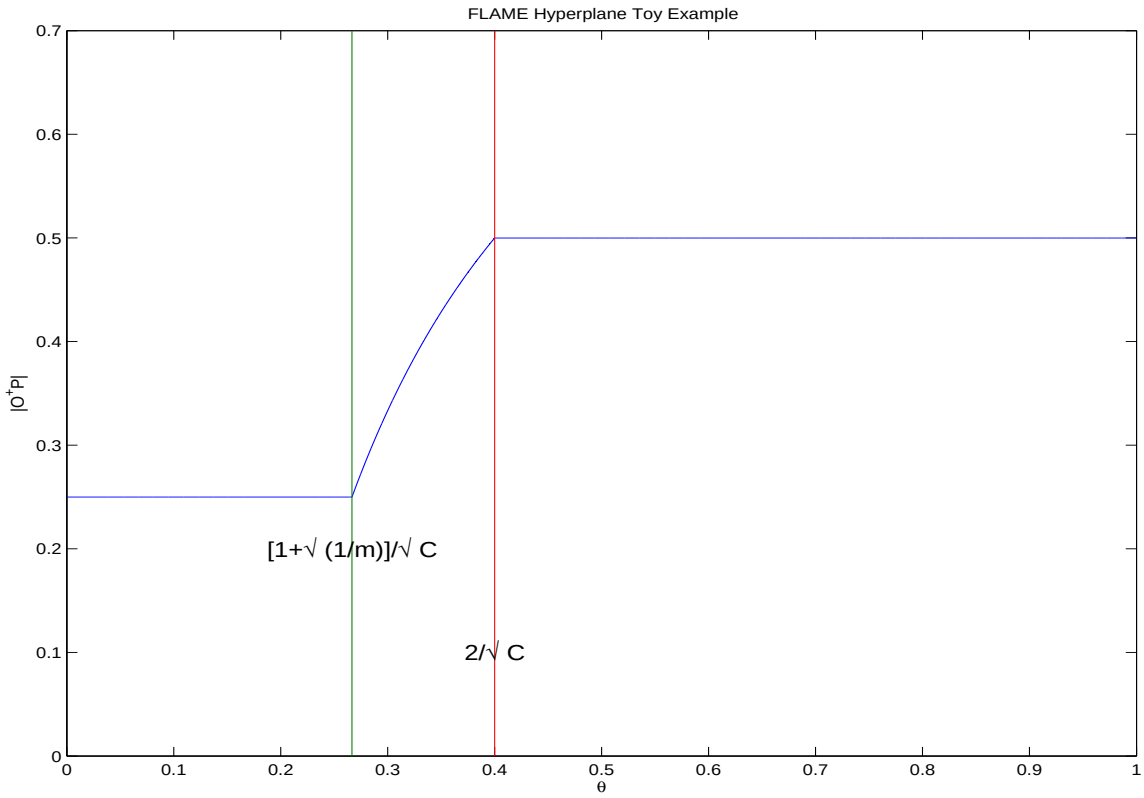

Figure A.1: A 1D toy example with $n_+ < n_-$ and $m = 9$ is used to mimic the d-asymptotic situation. The length of the line segment O^+O^- equals 1. As θ increases, the FLAME hyperplane stands still ($|O^+P|$ unchanged); when $\theta > \sqrt{1+(1/m)}/\sqrt{C}$, $|O^+P|$ increases, which means the hyperplane moves towards the negative class, until $\theta = 2/\sqrt{C}$, after which the hyperplane remains at the midpoint of O^+O^-.

representation above, in the d-asymptotics, the discriminant directions for most classifiers are the same. Moreover, the projections of data points in the same class to O^+O^- are the same, which is the normal vector for the DWD, SVM and FLAME hyperplanes. Therefore, they all have data-piling in the d-asymptotics.

9. Derivation of the FLAME Hyperplane in d Asymptotics

The FLAME seeks to minimize

$$n_+ \left(1/a - \theta\sqrt{C}\right)_+ + n_- \left(1/b - \theta\sqrt{C}\right)_+ \tag{3}$$

$$\text{s.t. } a + b = \sqrt{d(\mu^2 + \sigma^2/n_+ + \tau^2/n_-)} = \nu\sqrt{d}. \tag{4}$$

When $\theta \in \left[0, (1+\sqrt{m^{-1}})/(\nu\sqrt{dC})\right)$, it is easy to verify that both $(1/a - \theta\sqrt{C})_+$ and $(1/b - \theta\sqrt{C})_+$ are positive and equal to $1/a - \theta\sqrt{C}$ and $1/b - \theta\sqrt{C}$. In this case, the optimal solutions to problem 3, a and b satisfy $a/b = (n_+/n_-)^{1/2} = \sqrt{m^{-1}}$. In particular, $a = \sqrt{m^{-1}}/(1+\sqrt{m^{-1}})\nu\sqrt{d}$ and $b = 1/(1+\sqrt{m^{-1}})\nu\sqrt{d}$.

When $\theta \in \left[(1+\sqrt{m^{-1}})/(\nu\sqrt{dC}), 2/(\nu\sqrt{dC})\right)$, $1/b - \theta\sqrt{C} < 0$ and the optimized solutions to (3) are $b = 1/(\theta\sqrt{C})$ and $a = \sqrt{d}\nu - b$. Note that $a > b$.

When $\theta \in \left[2/(\nu\sqrt{dC}), 1\right]$, $a = b = 0.5\sqrt{d}\nu$

10. Proof to Theorem 7

We only need to prove the sure classification for the second interval, *i.e.*,

$$\theta \in \left[(1+\sqrt{m^{-1}})/(\nu\sqrt{dC}), 2/(\nu\sqrt{dC})\right).$$

The proofs for the other two intervals are similar to those in Qiao et al. (2010). It was shown in Hall et al. (2005) and Qiao et al. (2010) that the length of the line segment O^+O^- is $\sqrt{d}\nu$ and that the distance between the projection (denoted as P') of a new data point from the $\mathcal{X}^+$-population onto O^+O^- and the centroid of the positive class O^+ is $(\sigma^2/n_+)/(\mu^2 + \tau^2/n_-)$ times of its distance to the centroid of the negative class O^-, *i.e.*, $\dfrac{|O^+P'|}{|O^-P'|} = (\sigma^2/n_+)/(\mu^2+\tau^2/n_-)$, where $|AB|$ is the length of the line segment connecting points A and B. Denote $|O^+P'|$ as a' and $|O^-P'|$ as b'. Because $a' + b' = \sqrt{d}\nu$, we must have $b' = \sqrt{d}(\mu^2 + \tau^2/n_-)/\nu$. In order for this new data point to be correctly classified to the positive class, P' has to be the same side as O^+ with respect to the intersection of the FLAME hyperplane and O^+O^-, that is,

$$\begin{aligned}
& b' > b \\
\Leftrightarrow \quad & \sqrt{d}\left(\mu^2 + \frac{\tau^2}{n_-}\right)/\nu > \frac{1}{\sqrt{C}\theta} \\
\Leftrightarrow \quad & \mu^2 + \frac{\tau^2}{n_-} > \frac{1}{\sqrt{dC}\theta}\nu \\
\Leftrightarrow \quad & \nu^2 - \frac{\sigma^2}{n_+} > \frac{1}{\sqrt{dC}\theta}\nu \\
\Leftrightarrow \quad & \nu^2 - \frac{1}{\sqrt{dC}\theta}\nu - \frac{\sigma^2}{n_+} > 0 \\
\Leftrightarrow \quad & (\nu - \frac{1}{2\sqrt{dC}\theta})^2 - \frac{1}{4dC\theta^2} - \frac{\sigma^2}{n_+} > 0 \\
\Leftarrow \quad & \nu > \sqrt{\frac{1}{4dC\theta^2} + \frac{\sigma^2}{n_+}} + \frac{1}{2\sqrt{dC}\theta}
\end{aligned}$$

$$\Leftrightarrow \qquad \mu^2 > \left[\sqrt{\frac{1}{4dC\theta^2} + \frac{\sigma^2}{n_+}} + \frac{1}{2\sqrt{dC}\theta}\right]^2 - \frac{\sigma^2}{n_+} - \frac{\tau^2}{n_-}$$

$$\Leftrightarrow \qquad \mu^2 > T - \frac{\tau^2}{n_-}.$$

We now assume that P' is the projection of a new data point from the $\mathcal{X}^-$-population. In this situation, it can be shown that $a'/b' = (\mu^2 + \sigma^2/n_+)/(\tau^2/n_-)$ and thus $b' = \sqrt{d}\dfrac{\tau^2/n_-}{\nu}$. To correctly classify this new data point, we only need to have $b' < b$. That is,

$$\sqrt{d}\frac{\tau^2/n_-}{\nu} < b = \frac{1}{\theta\sqrt{C}}$$

$$\Leftrightarrow \qquad \frac{\tau^2}{n_-} < \frac{1}{\theta\sqrt{dC}}\sqrt{\mu^2 + \tau^2/n_- + \sigma^2/n_+}$$

We only need to show that $\dfrac{\tau^2}{n_-} < \dfrac{1}{\theta\sqrt{dC}}\sqrt{\tau^2/n_- + \sigma^2/n_+}$. Let $q^2 = \tau^2/n_- + \sigma^2/n_+$. We need to show that

$$q^2 - \frac{\sigma^2}{n_+} < \frac{1}{\theta\sqrt{dC}}q$$

$$\Leftrightarrow \qquad \left(q - \frac{1}{2\theta\sqrt{dC}}\right)^2 - \frac{1}{4\theta^2 dC} - \frac{\sigma^2}{n_+} < 0$$

$$\Leftarrow \qquad q < \sqrt{\frac{1}{4\theta^2 dC} + \frac{\sigma^2}{n_+}} + \frac{1}{2\theta\sqrt{dC}}$$

$$\Leftrightarrow \qquad \frac{\tau^2}{n_-} < \left[\sqrt{\frac{1}{4\theta^2 dC} + \frac{\sigma^2}{n_+}} + \frac{1}{2\theta\sqrt{dC}}\right]^2 - \frac{\sigma^2}{n_+}$$

$$\Leftrightarrow \qquad \frac{\tau^2}{n_-} < T.$$

The last inequality is the condition stipulated in the theorem. ■

11. Additional Figures

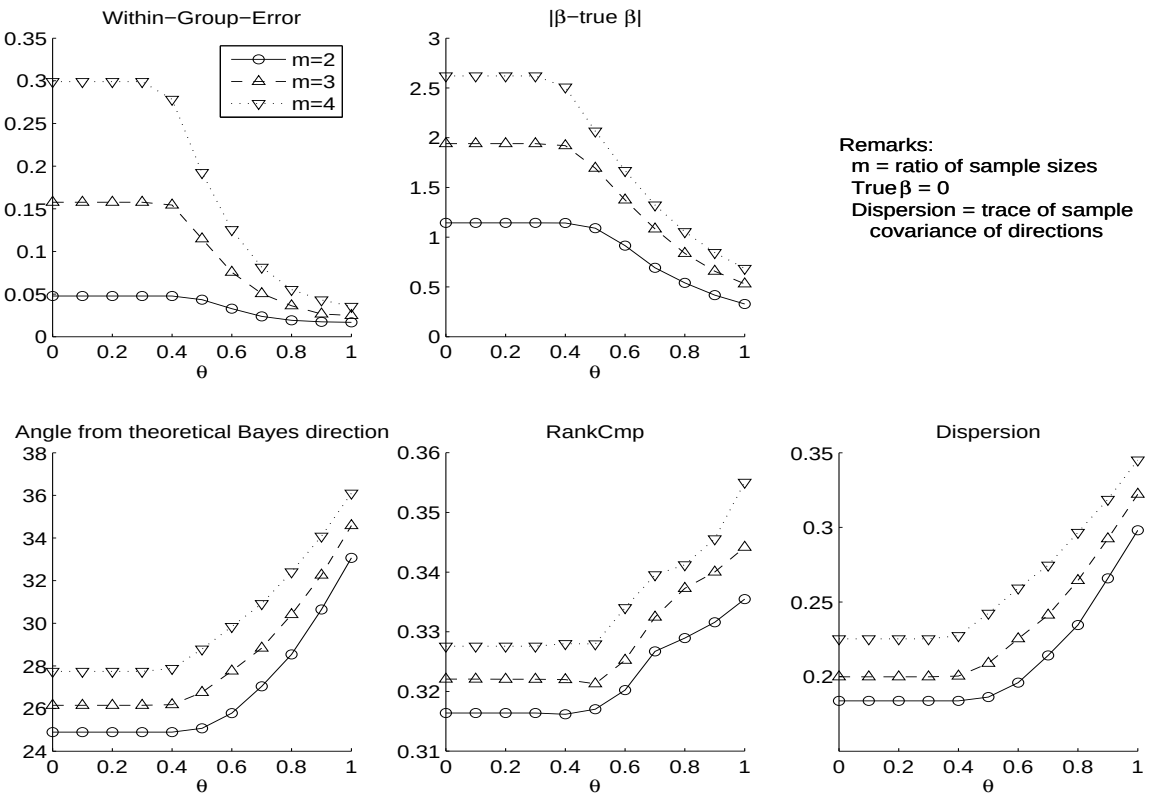

Figure A.2: Independent example. It can be seen that with FLAME turns from DWD to SVM (θ from 0 to 1), the within-class error decreases (top-left), thanks to the more accurate estimate of the intercept term (top-middle). On the other hand, this comes at the cost of larger deviation from the Bayes direction (bottom-left), incorrect rank of the importance of the variables (bottom-middle) and larger stochastic variability of the estimation directions (bottom-right).

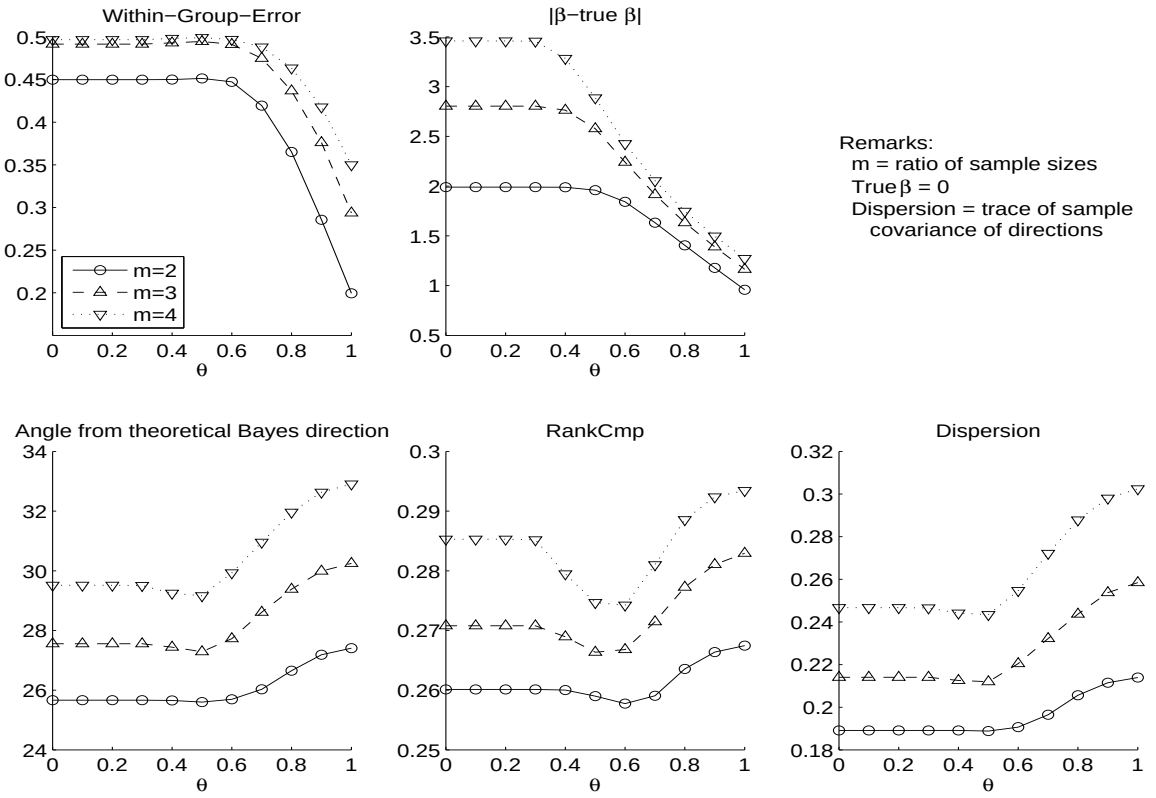

Figure A.3: Block interchangeable example. It can be seen that with FLAME turns from DWD to SVM (θ from 0 to 1), the within-class error decreases (top-left), thanks to the more accurate estimate of the intercept term (top-middle). On the other hand, this comes at the cost of larger deviation from the Bayes direction (bottom-left), incorrect rank of the importance of the variables (bottom-middle) and larger stochastic variability of the estimation directions (bottom-right). The RankComp measure in this example is an exception, in the sense that it decreases first then increases instead of monotonically increases.

References

Bartlett, P., Jordan, M., and McAuliffe, J. (2006), "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, 101, 138–156.

Hall, P., Marron, J. S., and Neeman, A. (2005), "Geometric Representation of High Dimension, Low Sample Size Data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 427–444.

Koo, J., Lee, Y., Kim, Y., and Park, C. (2008), "A Bahadur Representation of the Linear Support Vector Machine," *Journal of Machine Learning Research*, 9, 1343–1368.

Marron, J. S., Todd, M., and Ahn, J. (2007), "Distance-Weighted Discrimination," *Journal of the American Statistical Association*, 102, 1267–1271.

Owen, A. (2007), "Infinitely Imbalanced Logistic Regression," *Journal of Machine Learning Research*, 8, 761–773.

Qiao, X., Zhang, H., Liu, Y., Todd, M., and Marron, J. S. (2010), "Weighted Distance Weighted Discrimination and its Asymptotic Properties," *Journal of the American Statistical Association*, 105, 401–414.

Toh, K., Todd, M., and Tütüncü, R. (1999), "SDPT3–a MATLAB Software Package for Semidefinite Programming, Version 1.3," *Optimization Methods and Software*, 11, 545–581.

Tütüncü, R., Toh, K., and Todd, M. (2003), "Solving Semidefinite-Quadratic-Linear Programs Using SDPT3," *Mathematical Programming*, 95, 189–217.