2.1 $K = X \cdot X^T$

$= (\langle x_i \cdot x_j \rangle)_{ij}$.

$\|x_i - x_j\| = \sqrt{\langle x_i - x_j, x_i - x_j \rangle}$

$= \sqrt{\|x_i\|^2 + \|x_j\|^2 - 2\langle x_i, x_j \rangle}$

If we know K, we have $\begin{cases} \|x_i\|^2 = \langle x_i, x_i \rangle, \\ \|x_j\|^2 = \langle x_j, x_j \rangle \\ \langle x_i, x_j \rangle \end{cases}$

Hence, the distance between x_i and x_j and their norm can be given.

3.1 $J(w) = (Xw - y)^T \cdot (Xw - y) + \lambda w^T \cdot w$

$= w^T X^T X w - 2 w^T X^T y + \lambda w^T w + y^T y$

$\frac{d}{dw} J(w) = 2(X^T X w - X^T y + \lambda w)$

$(X^T X + \lambda I) w^* = X^T y$, $w^* = (X^T X + \lambda I)^{-1} X^T y$.

$X^T X + \lambda I$ must be positive definite when $\lambda > 0$.

So it's invertible.

3.2 $X^T \cdot X \cdot X^T\alpha + \lambda X^T\alpha = X^T y$ if $X^T\alpha = w^*$

$X^T(X \cdot X^T + \lambda I)\alpha = X^T y$.

$\alpha = (X \cdot X^T + \lambda I)^{-1} y$. can hold this equation.

Meanwhile, $X \cdot X^T + \lambda I$ is invertible since it's positive definitee.

So $w^* = X^T\alpha$, $\alpha = (X \cdot X^T + \lambda I)^{-1} y$.

3.3 $w = X^T \cdot \alpha$

$= (x_1, \cdots, x_n) \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$

$= \sum_{i=1}^{n} \alpha_i \cdot x_i$

w spans all x_i so it is said to be in the span of data.

3.4. Proved in 3.3.

3.5 $\cdot Xw^* = X \cdot X^T \cdot \alpha$

$= K \cdot \alpha$.

$= K(K + \lambda I)^{-1} y$.

3.6 $f(x) = x^T \cdot w^*$

$= x^T \cdot X^T \cdot \alpha$

$= K_x^T \cdot (K + \lambda I)^{-1} y$.

$K_x = x^T \cdot X^T$

$= x^T(x_1, \cdots, x_n)$

$= (x^T x_1, \cdots x^T x_n)$

4.1 $g_i(w) = \partial \frac{\lambda}{2}\|w\|_2^2 + v_i(w)$ [addicivity]

$= \lambda w + v_i(w)$

4.2 $E\, g_i(w) = \lambda w + \frac{1}{n}\sum_{i=1}^{n} v_i(w)$.

$\partial J(w) = \partial \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \partial l_i(w)$.

Because. $v_i(w)$ is a subgradient of $l_i(w)$

then $\lambda w + \frac{1}{n}\sum_{i=1}^{n} v_i(w)$ is a subgradient of $J(w)$.

Namely. $E[g_i(w)] \in \partial J(w)$

4.3. $w^{(t+1)} = w^{(t)} - \eta^{(t)} g_i(w^{(t)})$

$= w^{(t)} - \frac{1}{\lambda t}(\lambda w^{(t)} + v^{(t)})$

~~$= \frac{\lambda t-1}{\lambda t} w^{(t)} - \frac{1}{\lambda t} v^{(t)}$~~

$= \frac{t-1}{t} w^{(t)} - \frac{1}{\lambda t} v^{(t)}$

~~$= \frac{t-1}{t}\left(\frac{t-2}{t-1} w^{(t-1)} - \frac{1}{\lambda(t-1)} v^{(t-1)}\right) - \frac{1}{\lambda t} v^{(t)}$~~

~~$= \frac{t-2}{t-1} w^{(t+1)} - \frac{1}{\lambda t}\sum$~~

If $w^{(t)} = -\frac{1}{\lambda(t-1)}\sum_{\tau=1}^{t-1} v^{(\tau)}$, then

For $t=1$, $w^{(2)} = -\frac{1}{\lambda} v^{(1)}$, $w^{(t+1)} = w^{(t)} = \frac{t-1}{t}\left(-\frac{1}{\lambda(t-1)}\sum_{\tau=1}^{t-1} v^{(\tau)}\right) - \frac{1}{\lambda t} v^{(t)}$

$w^{(t+1)} = -\frac{1}{\lambda t}\sum_{\tau=1}^{t} v^{(\tau)} = \frac{t-1}{t}\left(-\frac{1}{\lambda(t-1)}\sum_{\tau=1}^{t-1} v^{(\tau)}\right) - \frac{1}{\lambda t} v^{(t)} = -\frac{1}{\lambda t}\sum_{\tau=1}^{t} v^{(t)}$

Inducively,

~~By induction.~~

~~$= \frac{t-1}{t} w^{(t)} - \frac{1}{\lambda t} v^{(t)}$~~

we have $w^{(t+1)} = \frac{1}{\lambda t}\sum v^{(\tau)}$

(a) ~~If $t=1$, $V_1(w) = y_i \langle w, x_i \rangle$~~

If $l_i(w) = \max\{0, 1 - y_i \langle w, x_i \rangle\}$

then $V_i(w) = \begin{cases} 0, & \text{else.} \\ -y_i x_i & y_i \langle w, x_i \rangle < 1 \end{cases}$ ②

Let $\theta^{(t+1)} = \sum_{\tau=1}^{t} V^{(\tau)}(w^{(\tau)})$, $w^{(t+1)} = -\frac{1}{\lambda t}\theta^{(t+1)}$.

$\theta^{(t+2)} = \theta^{(t+1)} + V^{(t+1)}(w^{(t+1)})$ $= -\frac{1}{\lambda t}\sum_{\tau=1}^{t} V^{(\tau)}(w^{(\tau)})$

$= \begin{cases} \theta^{(t+1)}, & \text{else.} \\ \theta^{(t+1)} - y^{(t+1)} x^{(t+1)}, & y^{(t+1)} \langle w^{(t+1)}, x^{(t+1)} \rangle < 1. \end{cases}$

$w^{(t+2)} = -\frac{1}{\lambda(t+1)}\theta^{(t+2)}$ ~~Ps~~ There is no redundant update compared with iteratively update.

The update rule showed above is precisely what's described in Algorithm 1.

~~P.S.: We can simply update $w^{(t)}$ as the iterative formula:~~

No need to maintain a variable for w cuz it can be computed by t and θ

~~$w^{(t+1)} = \frac{t-1}{t} w^{(t)} - \frac{1}{\lambda t} V^{(t)}$. No need to maintain a variable for θ. But if the number of parameters needed to update is increases, maintaining such variables is inevitable.~~

5.1 $y_j \langle w^{(t)}, x_j \rangle = \sum_{i=1}^{n} y_j \cdot \alpha_i \langle x_i, x_j \rangle$

$= y_j \cdot K_{j,:} \cdot \alpha^{(t)}$

5.2 If the point of (x_j, y_j) has no margin violation.

then $w^{(t+1)} = \frac{t-1}{t} w^{(t)}$, so $\alpha^{(t+1)} = \frac{t-1}{t} \alpha^{(t)}$

5.3 If the point of (x_j, y_j) has margin violation

$w^{(t+1)} = \frac{t-1}{t} w^{(t)} + \frac{1}{\lambda t} y_j x_j$.

So $\alpha^{(t+1)} = \frac{t-1}{t} \alpha^{(t)} + \frac{1}{\lambda t} y_j \cdot e_j$.

Pseudocode should be:

While convergence conditions don't satisfy:

For j in permutation series:

if $y_j \cdot K_{j,:} \cdot \alpha^{(t)} < 1$:

$\alpha^{(t+1)} = \frac{t-1}{t} \alpha^{(t)} + \frac{1}{\lambda t} y_j \cdot e_j$

else:

$\alpha^{(t+1)} = \frac{t-1}{t} \alpha^{(t)}$.

$t = t+1$

5.4 Let $\beta^{(t+1)} = \begin{cases} \beta^{(t)} + y^{(t)} \cdot e^{(t)}, & \text{if } y^{(t)} \cdot K^{(t),\cdot} \cdot \alpha^{(t)} < 1 \\ \beta^{(t)}, & \text{else.} \end{cases}$, ~~$\alpha^{(t+1)} = \frac{1}{\lambda t}\beta^{(t+1)}$~~

then $\alpha^{(t+1)} = \frac{1}{\lambda t}\beta^{(t+1)}$.

① For details, $\alpha^{(t+1)} = \frac{1}{\lambda t}\sum_{\tau=1}^{t} y^{(\tau)} \cdot v^{(\tau)}$, $v^{(\tau)} = \begin{cases} e^{(\tau)}, & \text{if } 1 - y^{(\tau)} \cdot K^{(\tau),\cdot} \cdot \alpha^{(\tau)} > 0 \\ 0, & \text{else.} \end{cases}$

by induction.

Let $\beta^{(t+1)} = \sum_{\tau=1}^{t} y^{(\tau)} \cdot v^{(\tau)} = \begin{cases} \beta^{(t)} + y^{(t)} \cdot e^{(t)}, & \text{if } 1 - y^{(t)} \cdot K^{(t),\cdot} \cdot \alpha^{(t)} > 0. \\ \beta^{(t)}, & \text{else} \end{cases}$

then $\alpha^{(t+1)} = \frac{1}{\lambda t}\beta^{(t+1)}$

② OR. Let $\beta^{(t+1)} = \lambda t \cdot \alpha^{(t+1)}$, then $\begin{cases} \frac{\beta^{(t+1)}}{\lambda t} = \frac{\beta^{(t)}}{\lambda t} + \frac{e^{(t)} \cdot y^{(t)}}{\lambda t}, & \text{if } 1 - \cdots > 0 \\ \frac{\beta^{(t+1)}}{\lambda t} = \frac{\beta^{(t)}}{\lambda t}, & \text{else.} \end{cases}$

It can be ~~simplified~~ simplified as: $\begin{cases} \beta^{(t+1)} = \beta^{(t)} + e^{(t)} \cdot y^{(t)}, & \text{if } \cdots \\ \beta^{(t+1)} = \beta^{(t)}, & \text{else.} \end{cases}$

For ① and ②:

No need to update ~~when~~ if no margin violation.

For ②:

We can easily find that $\beta^{(t+1)} = \sum_{\tau=1}^{t} y^{(\tau)} \cdot v^{(\tau)}$. v following the above definition. No need to induction.

implement

For ②: use this method in 4.3. $\alpha^{(t+1)} = \frac{t-1}{t}\alpha^{(t)} + \frac{1}{\lambda t}u^{(t)}$

$\Rightarrow \lambda t\, \alpha^{(t+1)} = (t-1)\alpha^{(t)} + u^{(t)}$.

Let $\beta^{(t+1)} = \lambda(t-1)\alpha^{(t)}$.

we have $\beta^{(t+1)} = \beta^{(t)} + u^{(t)}$!

then $\beta^{(t+1)} = \sum_{\tau=1}^{t} u^{(\tau)}$ ←

6.1 No problems.

6.2~6.4 Programming questions.

7.1 $x \in \mathcal{H}$ and m_0 is x's projection onto subspace M,

1. $$\|x\|^2 = \|m_0\|^2 + \|x - m_0\|^2$$

If $\|x\| = \|m_0\|$, then $\|x - m_0\|^2 = 0$. so $x = m_0$.

7.2 $J(w) = R(\sqrt{\langle w, w \rangle}) + L(\langle w, \varphi(x_1) \rangle, \cdots, \langle w, \varphi(x_n) \rangle)$,

$R(\cdot)$ is strictly increasing.

Consider M is a subspace of $\mathcal{H}$. and its basis is an linearly independent group of $\varphi(x_1), \cdots, \varphi(x_n)$.

For any $w \in \mathcal{H}$. $w = \sum_{i=1}^{n} \alpha_i \varphi(x_i) + r$, $r \perp M$. Denote $\sum_{i=1}^{n} \alpha_i \varphi(x_i) = m$. and $r \perp m$ either.

$J(w) = R(\sqrt{\langle w, w \rangle}) + L(\langle w, \varphi(x_1) \rangle, \cdots, \langle w, \varphi(x_n) \rangle)$

$= R(\sqrt{\langle m, m \rangle + \langle r, r \rangle}) + L(\langle m + r, \varphi(x_1) \rangle, \cdots, \langle m + r, \varphi(x_n) \rangle)$

$= R(\sqrt{\langle m, m \rangle + \langle r, r \rangle}) + L(\langle m, \varphi(x_1) \rangle, \cdots, \langle m, \varphi(x_n) \rangle)$

$\geq J(m)$.

So any minimizer of $J(w)$ has the form of $\sum_{i=1}^{n} \alpha_i \varphi(x_i)$.

7.3. $J[\alpha x+(1-\alpha)y] = R(\sqrt{\langle \alpha x+(1-\alpha)y, \alpha x+(1-\alpha)y\rangle}) + L(\langle \alpha x+(1-\alpha)y, \psi(x_1)\rangle, \cdots \langle \alpha x+(1-\alpha)y, \psi(x_N)\rangle)$

$= R(\sqrt{\alpha^2\langle x,x\rangle + 2(1-\alpha)\alpha\langle x,y\rangle + (1-\alpha)^2\langle y,y\rangle}) +$

$L[\alpha\langle x,\psi(x_1)\rangle + (1-\alpha)\langle y,\psi(x_1)\rangle, \cdots, \alpha\langle x,\psi(x_N)\rangle + (1-\alpha)\langle y,\psi(x_N)\rangle]$

$\alpha J(x) + (1-\alpha)J(y) = \alpha R(\sqrt{\langle x,x\rangle}) + (1-\alpha)R(\sqrt{\langle y,y\rangle})$

$+ \alpha L(\langle x,\psi(x_1)\rangle, \cdots, \langle x,\psi(x_N)\rangle)$

$+ (1-\alpha)L(\langle y,\psi(x_1)\rangle, \cdots, \langle y,\psi(x_N)\rangle)$

$\alpha R(\sqrt{\langle x,x\rangle}) + (1-\alpha)R(\sqrt{\langle y,y\rangle}) \geq R[\alpha\sqrt{\langle x,x\rangle} + (1-\alpha)\sqrt{\langle y,y\rangle}]$

$[\alpha\sqrt{\langle x,x\rangle} + (1-\alpha)\sqrt{\langle y,y\rangle}]^2 = \alpha^2\langle x,x\rangle + (1-\alpha)^2\langle y,y\rangle + 2\alpha(1-\alpha)\sqrt{\langle x,x\rangle\langle y,y\rangle}$

$\geq \alpha^2\langle x,x\rangle + (1-\alpha)\langle y,y\rangle + 2\alpha(1-\alpha)\langle x,y\rangle$

Hence, $R[\alpha\sqrt{\langle x,x\rangle} + (1-\alpha)\sqrt{\langle y,y\rangle}] \geq R(\sqrt{\alpha^2\langle x,x\rangle + 2(1-\alpha)\alpha\langle x,y\rangle + (1-\alpha)^2\langle y,y\rangle})$.

$\alpha J(x) + (1-\alpha)J(y) \geq J[\alpha x + (1-\alpha)y]$

So $J(\cdot)$ is convex function

8.1.1. For Tikhonov regularization, if:

$$f^* \in \underset{f \in F}{\operatorname{argmin}} [\phi(f) + \lambda \Omega(f)]$$

let $r = \Omega(f^*)$, its Ivanov form is:

$$\underset{\Omega(f) \le \Omega(f^*)}{\operatorname{argmin}} \phi(f).$$

If $f^* \notin \underset{\Omega(f) \le \Omega(f^*)}{\operatorname{argmin}} \phi(f)$, there will be f' satisfying $\phi(f') < \phi(f^*)$.

$$\phi(f') + \lambda \Omega(f') < \phi(f^*) + \lambda \Omega(f^*)$$

so $f^* \notin \underset{f \in F}{\operatorname{argmin}} [\phi(f) + \lambda \Omega(f)]$, which contradicts the former assumption.

Hence, $f^* \in \underset{\Omega(f) \le \Omega(f^*)}{\operatorname{argmin}} \phi(f)$.

8.2.1. The Lagrangian function for Ivanov regularization is:

$$L(w, \lambda) = \phi(w) + \lambda [\Omega(w) - r]$$

8.2.2 The dual optimization problem is:

$$\max_{\lambda} \min_{w} \phi(w) + \lambda[\Omega(w) - r]$$

8.2.3. Let $g(\lambda) = \min_{w} \phi(w) + \lambda[\Omega(w) - r]$

$$\ell(w) = \max_{\lambda} \phi(w) + \lambda[\Omega(w) - r]$$

If $\lambda^* \in \arg\max_{\lambda \geq 0} g(\lambda)$, $g(\lambda^*) \leq \phi(w) + \lambda^*[\Omega(w) - r]$

According to strong duality, $g(\lambda^*) = \ell(w^*) = \phi(w^*)$

$$\phi(w^*) + \lambda^*[\Omega(w^*) - r] = \phi(w^*) = g(\lambda^*)$$

Hence, $w^* \in \arg\min_{w} \phi(w) + \lambda^*[\Omega(w) - r]$

Omit the constant term $-\lambda^* r$, $w^* \in \arg\min_{w} \phi(w) + \lambda^* \Omega(w)$

8.2.4. $g(\lambda) \leq \phi(w) + \lambda[\Omega(w) - r]$, $g(0) \leq \phi(w)$

~~If $0 = \lambda^* \in \arg\min_{\lambda} g(\lambda)$, $g(0)$~~

If $\inf \phi(w) < \inf_{\Omega(w) \leq r} \phi(w)$, there must be a $w' \in \arg\inf_{w} \phi(w)$,

$\Omega(w') > r$.

$g(0) \leq \phi(w') < \phi(w^*)$, it contradicts with strong duality if

$0 \in \arg\max_{\lambda \geq 0} g(\lambda)$.

Hence, $\lambda^* > 0$. Because of complementary slackness. $\|w^*\| = r$.