1. SVM, $y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i$, $\varepsilon_i \geq 0$.
Slack variables, distance that need to move bad point.

2. Convex Set D, $x \in D$, $y \in D$,
$\lambda x + (1-\lambda) y \in D$, $\lambda \in [0,1]$
Convex Function, with domain D
$f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2)$
$f(y) \geq f(x) + \nabla f_x (y - x)$.

3. $f(\vec{x})$ vector valued, $= [x_1^2, x_2^2 \cdots x_n^2]^T$

$\frac{\partial f}{\partial \vec{x}} = \nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$ $\quad \nabla_x^2 f = \frac{\partial f_i}{\partial x_j}$ ← column ← row

$f(\vec{x}) = \begin{bmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \end{bmatrix}$, $\nabla_x f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} \\ \vdots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} \end{bmatrix}$

$\nabla_x^2 f$ is p.s.d. $\Leftrightarrow$ f is convex
positive definite (p.d) $\Leftrightarrow$ strictly convex

4. A is p.s.d if $x^T A x \geq 0$ for all x
$x^T A x = x_1^2 a_{11} + x_1 x_2 a_{12} + x_2 x_1 a_{21} + x_2^2 a_{22}$
A is p.d if $x^T A x > 0$ for all x

5. A is p.s.d if eig value all ≥ 0
A is p.s.d if $A = B^T B$

6. If A p.s.d, then $\|Ax - b\|_2^2$ convex

7. If $G(x)$ is convex, F is matrix, then $G(Fx)$ convex

8. $\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$

$\frac{\partial Ax}{\partial x} = A^T$

$\frac{\partial \|Ax - b\|_2^2}{\partial x} = \frac{\partial (Ax-b)^T (Ax-b)}{\partial x} = 2A^T(Ax - b)$

$\frac{\partial (x-s)^T A (x-s)}{\partial x} = (A + A^T)(x - s)$

$\sigma(z) = 1/1+e^{-z}$

$\frac{\partial \sigma}{\partial z} = \sigma(z)/(1 - \sigma(z))$

$(f \cdot g)' = f' \cdot g + g' \cdot f$

$(\frac{f}{g})' = \frac{f' \cdot g - g' \cdot f}{g^2}$

$(e^x)' = e^x$
$(\log x)' = \frac{1}{x}$

unconstrained optimization, check what happened as some components go to $\pm\infty$, check bounding condition if some components are bounded.

9. Gradient Descent $\vec{\theta}^{(i+1)} \leftarrow \vec{\theta}^{(i)} - \eta(i) \nabla f(\vec{\theta}^{i})$
learning rate $\sum \eta(i) = \infty$, but $\sum \eta(i)^2 < \infty$. e.g. $\eta(i) = \frac{c}{i}$
- Momentum: $A_i = \vec{\theta}^{i+1} - \vec{\theta}^{i} = \eta \nabla f + \alpha A_{i-1}$
- Stochastic gradient descent.

10. Chain Rule: $f(x) = g(h(x))$, $\frac{\partial f}{\partial x} = \frac{\partial h(x)}{\partial x} \cdot \frac{\partial g(x)}{\partial h}$

$f(x) = g(h(\vec{x})) = g\begin{pmatrix} h_1(\vec{x}) \\ \vdots \\ h_k(\vec{x}) \end{pmatrix}$ $\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}, \frac{\partial h_2}{\partial x_1} \cdots \frac{\partial h_k}{\partial x_1} \\ \vdots \\ \frac{\partial h_1}{\partial x_n}, \frac{\partial h_2}{\partial x_n} \cdots \frac{\partial h_k}{\partial x_n} \end{bmatrix} \begin{bmatrix} \frac{\partial g}{\partial h_1} \\ \vdots \\ \frac{\partial g}{\partial h_k} \end{bmatrix}$

Dual = $\max_{\alpha} \min_{x} D(x)$.
for inequality, $\vec{\alpha} \geq 0$.

11. KKT condition: $\min f(x)$, subject to $g_i(x) \leq 0$, $h_j(x) = 0$.
① $\frac{\partial L(x^*, \alpha^*, \beta^*)}{\partial x} = 0$ ② $h(x^*) = 0$. ③ $\alpha_i^* g_i(x^*) = 0$.
Note in ①, in augmented form, do not take augmented.
④ $g_i(x^*) \leq 0$. ⑤ $\alpha_i^* \geq 0$.

12. Check if Dual applied. ④ slater condition, exist x, $h(x)=0$, $g(x) < 0$
① $f(x)$ is convex ② (inequality) $g(x)$ is convex ③ $h(x)$ is linear

13. ADMM. P1
$\underset{x, z}{\operatorname{argmin}} \underset{①}{\frac{1}{2}(z-u)^2} + \underset{②}{\lambda|x| + \alpha(x-z)}$

original problem
$\operatorname{argmin} \frac{1}{2}\|x - u\|_2^2 + \lambda \|x\|_1$

$(\lambda > 0)$ regularization constant.
② $\min \lambda|x| + \alpha x$,
if $\alpha > \lambda$, then $-\infty$, because set $x = -\infty$
$\alpha = \lambda$, then 0, because $x = 0$ or $x < 0$
$\alpha = -\lambda$, then 0, because $x = 0$ or $x > 0$
$-\lambda < \alpha < \lambda$, then 0, because $x = 0$.
$\alpha < -\lambda$, $-\infty$, because $x = +\infty$
$\Rightarrow \begin{cases} 0 & \text{if } |\alpha| \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$

So $D(\alpha) = \begin{cases} -\frac{1}{2}\alpha^2 - \alpha u & \text{if } |\alpha| \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$

$\max -\frac{1}{2}\alpha^2 - \alpha u$ s.t. $|\alpha| \leq \lambda$
$\frac{\partial D}{\partial \alpha} = -\alpha - u \Rightarrow \alpha = \begin{cases} -u, & |u| \leq \lambda \\ -\lambda, & u \geq \lambda \\ \lambda, & -u \leq \lambda \end{cases}$

$\max D(\alpha)$
① $\Rightarrow x = z = u + \alpha \Rightarrow$ so $x = S_\lambda(u) = \begin{cases} u - \lambda, & u \geq \lambda \\ 0, & \text{if } |u| \leq \lambda \\ u + \lambda, & \text{if } u \leq -\lambda \end{cases}$

14. P2: $\min \frac{1}{2}\|\vec{x} - \vec{u}\|_2^2 + \lambda \|\vec{x}\|_1$, x_i are independent.
$= \sum_{i=1}^{K} \left[\frac{1}{2}(x_i - u_i)^2 + \lambda |x_i| \right]$, so answer $\vec{y}$, $y_i = S_\lambda(u_i)$.

15. P3. $\underset{x}{\operatorname{argmin}} |x - u| + \lambda |x|$. $\Rightarrow D(\alpha) = \min_{x,z} |z - u| + \lambda|x| + \alpha(x - z)$

$D(\alpha) = \min \underset{①}{|z-u| + \alpha z} + \min \underset{②}{\lambda |x| + \alpha x}$
$= \begin{cases} -\alpha u & \text{if } |\alpha| \leq 1 \\ -\infty & \text{otherwise} \end{cases} + \begin{cases} 0 & \text{if } |\alpha| \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$

$\begin{cases} -\alpha u, \text{ if } |\alpha| \leq \min(1, \lambda) \end{cases}$
if $u > 0$, $\alpha = -\min(1, \lambda)$
$u = 0$, $x = 0$
$u < 0$, $\alpha = \min(1, \lambda)$

$\min_z |z - u| - \alpha z$, set $y = z - u$.
$y' = -y$.
See 13. ②

Dual Solution $\begin{cases} \alpha = -\min(1, \lambda), & u > 0 \\ \alpha = \min(1, \lambda), & u < 0 \end{cases}$

If $|\alpha| > 1$, $z = \pm\infty$
$\alpha = 1$ $z \geq u$.
$\alpha = -1$, $z \leq u$.
$|\alpha| < 1$, $z = u$.

Suppose $u > 0$.
$\Rightarrow$ Case 1, $\lambda > 1$, so $\alpha = -1 > \lambda$, so $x = z \leq u$ from z, $x = 0$, from x sol
Case 2. $\lambda < 1$, so $\alpha = -\lambda$, $x = z = u$. from z sol
Case 3. $\lambda = 1$. so $\alpha = -1 = -\lambda$ $\left.\begin{matrix} x = z \leq u \text{ from } z \\ x \geq 0 \text{ from } x \end{matrix}\right\}$ given
$0 \leq x \leq u$

1. P4. $\arg\min_{\vec z} \frac12\|\vec x-\vec u\|_2^2+\lambda\|\vec x\|_\infty$

$\arg\min \frac12\|\vec z-\vec u\|_2^2+\lambda\|\vec x\|_\infty+\vec\alpha^T(\vec x-\vec z)$

$\min_z \frac12\|\vec z-\vec u\|_2^2-\vec\alpha^T\vec z+\min_x \lambda\|\vec x\|_\infty+\vec\alpha^T\vec x$

$\Downarrow$ $z=u+\alpha$ $\Downarrow$

$\min \lambda\|x\|_\infty+\vec a^T\vec x$

Change of vars. $\vec x=r\vec y$, $r=\|x\|_\infty$. y is vector with $\|\vec y\|_\infty=1$, so we $\min_{r,\vec y} \lambda r+r\vec a^T\vec y$.

Choosing a good $\vec y$, $\sum_{i=1}^K r\alpha_i y_i$. s.t. $|y_i|\le 1$, for all i

so $y=-1$ if $\alpha_i\ge 0 \Rightarrow \vec y=-\text{sign}(\vec\alpha)$,
+1 otherwise

so $\sum r\alpha_i y_i=-r\|\vec\alpha\|_1 \Rightarrow \min \lambda r-r\|\alpha\|_1$

solution $=\begin{cases}0 & \text{if } \lambda\ge\|\alpha\|_1\\ -\infty & \text{otherwise}\end{cases} \Rightarrow D(\alpha)=\frac12\|\vec\alpha\|_2^2-\vec\alpha^T(\vec u+\vec\alpha)$ s.t. $\|\alpha\|_1\le\lambda$.

2. P5. $\arg\min \frac12\|\vec x-\vec u\|_2^2+\lambda\|\vec x\|_2$

$\min_{\vec x,\vec z} \frac12\|\vec z-\vec u\|_2^2+\lambda\|\vec x\|_2+\alpha^T(\vec x-\vec z)$

$\min \frac12\|\vec z-\vec u\|_2^2-\alpha^T z+\min \lambda\|\vec x\|_2+\alpha\vec x$

$\vec z=\vec u+\vec\alpha$ $\Downarrow$

$r=\|x\|_2$. $\vec y$ unit vector, $\|y\|_2=1$.

$\min_{r,y} \lambda r+r a^T y$, $r\ge 0$, $\|y\|_2=1$.

$\vec y$ is in opposite direction of $\vec\alpha$. $y=-\frac{\alpha}{\|\alpha\|_2}$.

so. $\min_{r\ge 0} \lambda r-r\|\alpha\|_2$. so, $r=\begin{cases}0 & \text{if } \|\alpha\|_2\le\lambda\\ \infty & \text{otherwise.}\end{cases}$

$\Rightarrow$ Sol $=\begin{cases}0 & \text{if } \|\alpha\|_2\le\lambda\\ -\infty & \text{otherwise}\end{cases}$

so. $D(\vec\alpha)=-\frac12\|\vec\alpha\|_2^2-u^T\vec\alpha$ s.t. $\|\alpha\|_2\le\lambda$.

$\vec\alpha=r\vec y \Rightarrow \max_{r,y} -\frac12 r^2-ry^Tu$, s.t. $r\le\lambda$.

$\vec y=-\frac{u}{\|u\|_2} \Rightarrow \max_{0\le r\le\lambda} -\frac12 r^2+r\|u\|_2$

$\frac{\partial}{\partial r}=-r+\|u\|_2$, $r=\|u\|_2$, $\Rightarrow r=\begin{cases}\|u\|_2 & \text{if } \|u\|_2\le\lambda\\ \lambda, & \text{otherwise } (\|u\|_2>\lambda)\end{cases}$

So $\vec\alpha=r\vec y=\begin{cases}-u & \text{if } \|u\|_2\le\lambda.\\ -\lambda u/\|u\|_2 & \text{otherwise}\end{cases}$

$z=x=\vec u+\vec\alpha$ $\Downarrow$

$\vec x=\vec z=\begin{cases}0 & \text{if } \|u\|_2\le\lambda \Rightarrow \text{// switch that turn off x components}\\ (1-\lambda/\|u\|_2)u & \text{otherwise.}\end{cases}$

Partition x into groups: $[\underbrace{x_1, x_2, x_3}_{x_{G_1}}, \underbrace{x_4\, x_5\, x_6}_{x_{G_2}}, \underbrace{x_7\, x_8\, x_9}_{x_{G_3}}]$

$\arg\min \frac12\sum\|x_{G_i}-u_{G_i}\|_2^2+\sum\lambda_i\|x_{G_i}\|_2$

Solu: $x_{G_i}=\begin{cases}0 & \text{if } \|u_{G_i}\|\le\lambda_i\\ \left(1-\frac{\lambda_i}{\|u_{G_i}\|_2}\right)u_{G_i}\end{cases}$

3. ADMM $\min f(x)+g(z)+\frac\rho2\|Ax+Bz-c\|_2^2$ s.t. $Ax+Bz-c=0$

$x_{t+1}\leftarrow\arg\min f(x)+\frac\rho2\|Ax+Bz_t-c\|_2^2+\alpha_t^TAx$

$z_{t+1}\leftarrow\arg\min g(z)+\rho/2\|Ax_{t+1}+Bz-c\|_2^2+\alpha_t^TBz$

$\alpha_{t+1}\leftarrow\alpha_t+\rho(Ax_{t+1}+Bz_{t+1}-c)$.

KTT: $Ax^*+Bz^*-c=0$
$\vec 0\in\partial_x f(x^*)+A^T\alpha^*$
$\vec 0\in\partial_z g(z^*)+B^T\alpha^*$

Consider x update:

$0\in\partial_x f(x_{t+1})+\rho A^T(Ax_{t+1}+Bz_t-c)+A^T\alpha_t$

$0\in\partial_x f(x_{t+1})+A^T(\rho(Ax_{t+1}+Bz_t-c)+\alpha_t)$

replace z_t with z_{t+1}, this is $\alpha_{t+1}-\rho Bz_{t+1}+\rho Bz_t$:

$0\in\partial_x f(x_{t+1})+A^T(\alpha_{t+1}-\rho B(z_{t+1}-z_t))$.

$0\in\partial_x f(x_{t+1})+A^T\alpha_{t+1}-\rho A^TB(z_{t+1}-z_t)$

monitor convergence ① $\rho A^TB(z_{t+1}-z_t)\to 0$ small ρ better

② $Ax_{t+1}+Bz_{t+1}-c\to 0$ large ρ better

$u=(1/\rho)\cdot\alpha$. $x_{t+1}=\arg\min(f(x)+\frac\rho2\|Ax+Bz_t-c+u_t\|_2^2$

$z_{t+1}=\arg\min(g(z)+\frac\rho2\|Ax_{t+1}+Bz-c+u_t\|_2^2$, $u_{t+1}=u_t+(Ax_{t+1}+Bz_{t+1}-c)$

$f(x)=\mathbb{I}_C(x)$. $x^*=\Pi_C(v)$ | $f(x)=\lambda\|x\|_1$. $x^*=S_{\lambda/\rho}(v)=\begin{cases}v-\lambda/\rho\,\text{sign}(v), & |v|>\lambda/\rho\\ 0, & \text{if } |v|<\lambda/\rho\end{cases}$

when $x^*=\arg\min f(x)+\frac\rho2\|x-v\|_2^2$

$\min\|x\|_1+\mathbb{I}_C(z)$. $C=\{x\mid Ax-b=0\}$, s.t. $x=z$

4. $\min\|x\|_1$ subject to $Ax=b \Rightarrow$

$\min\|x\|_1+I_c(z)+y^T(x-z)+\frac\rho2\|x-z\|_2^2$. $u=1/\rho\cdot y$

$\rho u^T(x-z)+\rho/2\|x-z\|_2^2=\frac\rho2(\|x-z\|_2^2+2u^T(x-z)+\|u\|_2^2)-u^Tu$

$=\frac\rho2\|x-z+u\|_2^2-\frac\rho2\|u\|_2^2$. so: $\min\|x\|_1+I_C(z)+\frac\rho2\|x-z+u\|_2^2$

$x^{k+1}=\arg\min\|x\|_1+\frac\rho2\|x-z^k+u^k\|_2^2=S_{1/\rho}(z^k-u^k)$,

$z^{k+1}=\arg\min I_C(z)+\frac\rho2\|x^{k+1}-z+u^k\|_2^2=\Pi_C(x^{k+1}+u^k)$ 解释

$u^{k+1}=u^k+(x^{k+1}-z^{k+1})$

$I_c(x)=\begin{cases}0 & \text{if } x\in C\\ \infty, & x\notin C\end{cases}$

5. $\frac\rho2\|x-y+u\|_2^2+\frac\rho2\|x-z+v\|_2^2=\rho\|x-\frac{\mu+\eta}{2}\|$. $\mu=y-u$, $\eta=z-v$

$f(x)=\frac12(x-u)^2+\lambda|x|$.

$\partial f(x)=\begin{cases}\{x-u+\lambda\}, & \text{if } x>0,\\ [x-u-\lambda, x-u+\lambda]=[-u-\lambda,-u+\lambda] & \text{if } x=0,\\ \{x-u-\lambda\} & \text{if } x<0\end{cases}$

6. LASSO. $\min\frac12\|Ax-b\|_2^2+\lambda\|z\|_1$. $Fx-z=0$.

$L_\rho=\frac12\|Ax-b\|_2^2+\lambda\|z\|_1+\frac\rho2\|Fx-z+u\|_2^2$

$x^{k+1}=(A^TA+\rho F^TF)^{-1}(A^Tb+\rho F^T(z^k-u^k))$

$z^{k+1}=S_{\lambda/\rho}(Fx^{k+1}+u^k)$; $u^{k+1}=u^k+Fx^{k+1}-z^{k+1}$

7. subgradient of f at $\vec x$ is any vector $\vec v$, s.t. $f(\vec y)\ge f(\vec x)+(\vec y-\vec x)\cdot\vec v$ for all $\vec y$.

If f is convex and $\vec 0\in\partial f(x)$, then x is minimizer of f

8. SVM. $\min_w \lambda\sum_{i=1}^N(1-y_i(w^Tx_i+b))+\frac12\|w\|_2^2$, $A=\begin{bmatrix}-y_1x_1 & -y_1\\ \vdots & \\ -y_mx_m & -y_m\end{bmatrix}$, $v=\begin{bmatrix}w\\ b\end{bmatrix}\in\mathbb{R}^{d+1}$

$\Rightarrow \min\lambda\sum 1^T(A_iv_i+1)+\frac12\|z\|_2^2$. s.t. $v_i-z=0$

$\sum_{i=1}^B 1^T(A_iv_i+1)+\frac1{2\lambda}\|z\|_2^2+\sum_{i=1}^B\frac\rho2\|v_i-z+u_i\|_2^2$.

$v_i^{k+1}=\arg\min 1^T(A_iv_i+1)+\frac\rho2\|v_i-z^k+u_i\|_2^2$

$z^{k+1}=\arg\min\frac1{2\lambda}\|z\|_2^2+\frac\rho2\sum\|v_i^{k+1}-z+u_i\|_2^2$

$u_i^{k+1}=u_i^k+(v_i^{k+1}-z^{k+1})$

$\min_w\frac12\|w\|_2^2+C\sum_{i=1}^N\varepsilon_i$

s.t. $y_i(w^Tx_i+b)\ge 1-\varepsilon_i$.

$\forall i$. $\varepsilon_i\ge 0$

- Variables:

1. x. observed variables (like data)
2. z. missing stuff (random variables we do not observe)
3. θ parameters (not known and not random)

Model must specify $p(z|\theta)$, $p(x|z,\theta)$.

θ is usually split into 2 parts $p(z|\theta_1)$, $p(x|z,\theta_2)$

Goal: 1. Find value of θ so that model fits the data.

2. Find (approximate) posterior dist $p(z|x,\theta)$

both by maximizing likelihood $= p(x|\theta_1,\theta_2) = \sum_z p(x|z,\theta_2)p(z,\theta_1)$

$\Lambda(q,\theta) = \sum_z q(z)\cdot\log\frac{p(z,x|\theta)}{q(z)}$.

1) If we could fully optimize q, then q would be $p(\cdot|x,\theta)$. because Λ is maximized when $KL(q\|p(\cdot|x,\theta))$ is minimized.

2). If we can't compute $p(z|x,\theta)$, then $q(z)$ is an approximation $p(z|x,\theta)$.
or if $z=(z_1,z_2)$. then $q(z)=q_1(z_1)q_2(z_2)$.

Exponential dist. $y\sim f(y|\beta)=\beta e^{-\beta y}\, 1\{y\ge0\}$

$\int u\,dv = uv - \int v\,du$

mean $= 1/\beta$.

variance $= 1/\beta^2$

$$= \int_0^\infty \beta y e^{-\beta y}dy$$
$$= -ye^{-\beta y}\Big|_0^\infty - \int_0^\infty e^{-\beta y}dy$$

$y\sim$ Exponential(β). $x_1 - x_n$ independently from Gauss(y,σ^2).

x observed: $x_1..x_n$; z missing $= y$ | θ param: β, σ^2

In this example, missing variable is only sampled once.

Full data likelihood $p(x,z|\theta)=p(x|z,\theta)p(z|\theta)$

$$=\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-y)^2/2\sigma^2}\right)\beta e^{-y\beta}\, 1\{y\ge0\}. \quad (*)$$

$\Lambda(q,\theta)$ function is $\sum_z q(z)\log\frac{p(x,z|\theta)}{q(z)}$.

set $q(y)$ = Exponential(α) $=\alpha e^{-\alpha y}\,1\{y\ge0\}$

$$\Lambda(q,\theta)=\Lambda(q,\beta,\sigma^2)=\sum_z q(z)\log\frac{p(x,z|\theta)}{q(z)}$$
$$=\int_0^\infty \alpha e^{-\alpha y}\log\frac{(*)}{\alpha e^{-\alpha y}}dy$$
$$=\int_0^\infty \alpha e^{-\alpha y}\left[\left(\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - (x_i-y)^2/2\sigma^2\right) + \log\beta - y\beta - \log\alpha + \alpha y\right]dy$$
$$=-\frac{n}{2}\log(2\pi\sigma^2)+\log\beta-\log\alpha$$
$$+\int_0^\infty\left[\left(\sum_{i=1}^k -(x_i-y)^2/2\sigma^2\right) - y\beta+\alpha y\right]\alpha e^{-\alpha y}dy$$

$\int_0^\infty y\alpha e^{-\alpha y}dy$ = mean of exponential $=\frac1\alpha$.

$$\int_0^\infty (y-c)^2\alpha e^{-\alpha y}dy = E(y-c)^2 = E\left(y-\frac1\alpha+\frac1\alpha-c\right)^2$$
$$=E\left(y-\frac1\alpha\right)^2 + 2\underbrace{E\left[y-\frac1\alpha\right]}_{0}\left(\frac1\alpha-c\right)+\left(\frac1\alpha-c\right)^2$$

var $=1/\alpha^2$

$$=1/\alpha^2+\left(\frac1\alpha-c\right)^2$$

So $\Lambda(\alpha,\beta,\sigma^2) = -\frac n2\log(2\pi\sigma^2)+\log\beta-\log\alpha-\frac\beta\alpha+1 - \frac{1}{2\sigma^2}\sum_{i=1}^n\left[\frac{1}{\alpha^2}+\left(x_i-\frac1\alpha\right)^2\right]$

Project onto a convex set C: $\arg\min \frac12\|x-y\|_2^2$ s.t. $Ax=b$

project onto set $\{x|Ax=b\}$. $\Lambda=\frac12\|x-y\|_2^2+\alpha^T(Ax-b)$

dual problem $\arg\max_\alpha\left[\arg\min_x \Lambda(x,y)\right]$

$\arg\min_x \frac12\|x-y\|_2^2+\alpha^T(Ax-b)$ | gradient is column vector, α^TA is a row vector

$\frac{\partial}{\partial x}= x-y+A^T\alpha=0 \Rightarrow x=y-A^T\alpha$

plug in: $\frac12\|A^T\alpha\|_2^2+\alpha^TAy-\|A^T\alpha\|_2^2-\alpha^Tb$ | $\alpha=(AA^T)^{-1}(Ay-b)$

$=-\frac12\|A^T\alpha\|_2^2+\alpha^TAy-\alpha^Tb$, $\frac{\partial}{\partial\alpha}=-AA^T\alpha+Ay-b=0$

Since $x=y-A^T\alpha=y-A^T(AA^T)^{-1}(Ay-b)$.

special case A is a vector w^T. $x=y-\frac{w}{\|w\|_2}(yw-b)$

If $w\cdot y=b$ then $x=y$

(*) $\arg\min\frac12\|x-y\|_2^2$ s.t. $wx\le b \Rightarrow$ If $w\cdot y>b$, then x is a boundary (boundary is convex) $wx=b$

$\arg\min\frac12\|x-y\|_2^2$ s.t. $w_1x\le b_1$, $w_2x\le b_2$, $w_3x\le b_3$. now use ADMM.

$\arg\min\frac12\|x-y\|_2^2+I\{w_1z_1\le b_1\}+I\{w_2z_2\le b_2\}+I$.. s.t. $x=z$, $x=v$..

$x\leftarrow\arg\min\frac12\|x-y\|_2^2+\frac\rho2\|x-z+\alpha_1\|_2^2+\frac\rho2\|x-v+\alpha_2\|_2^2+\cdots$

$z\leftarrow\arg\min I(w_1z\le b)+\frac\rho2\|x-z+\alpha_1\|_2^2$ projection of $x+\alpha_1$ onto (*)

$p\sim$ Beta(a,b). density $f(p)=\frac{\Gamma(\alpha+b)}{\Gamma(a)\Gamma(b)}p^{\alpha-1}(1-p)^{b-1}\,1\{p\in[0,1]\}$

mean $=\frac{\alpha}{\alpha+b}$ variant $\frac{ab}{(a+b)^2(a+b+1)}$ | $p\sim$ Beta(α,b). $x\sim$ Bernoulli(n,p)

Joint: $p(x=k,p=p)=\binom nk p^k(1-p)^{n-k}\frac{\Gamma(\alpha+b)}{\Gamma(\alpha)\Gamma(b)}p^{\alpha-1}(1-p)^{b-1}$

$=\binom nk\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(k+a)\Gamma(n-k+b)}{\Gamma(n+a+b)}$, $p(p|k)=\frac{p(p,k)}{p(k)}=\frac{\Gamma(n+a+b)}{\Gamma(k+a)\Gamma(n-k+b)}p^{k+a-1}(1-p)^{n-k+b-1}$

= Beta(k+a, n-k+b)

$f(p)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}$.

$\int_0^1 f(p)dp=1 \Rightarrow \int_0^1 p^{\alpha-1}(1-p)^{b-1}dp=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

conjugate: $p(p|k)$ is the same dist as p | $KL(q\|p)=\sum_z q(z)\log\frac{q(z)}{p(z)}$.

$KL(q\|p)\ge0$., $KL(q\|q)=0$, $\arg\min_q KL(q\|p)=p$, $\arg\min_p KL(q\|p)=q$

#2. Fixed cluster centers $\mu_1 \dots \mu_K$ (unknown)

to generate a point x_j. ① pick a cluster i with prob π_i (multinomial$(1, \pi)$). ② $x_j \sim \text{Gauss}(\mu_i, 1)$

1. observed: $x_1 \dots x_n$. 2. missing: cluster of x_j, call it z_j.
3. unknown: $\pi(\pi_1, \dots \pi_K)$, $\mu_1 \dots \mu_K$.

$z_j = [0 \dots 1 \dots 0]$ 1 coord is 1. everything else is 0.

$z_j[i] = 1$ means x_j is in cluster i.

complete data likelihood function

$$p(x_1 z_1 x_2 z_2 \dots x_n z_n \mid \mu_1 \dots \mu_K, \pi)$$

$$= \prod_{j=1}^{n} p(x_j \mid z_j)\, p(z_j \mid \mu_1 \dots \mu_K, \pi)$$

$$= \prod_{j=1}^{n}$$ if cluster of point j is 1 then $\frac{1}{\sqrt{2\pi}} e^{-(x_j-\mu_1)^2/2}\pi_1$

——— 2. then $\frac{1}{\sqrt{2\pi}} e^{-(x_j-\mu_2)^2/2}\pi_2$

$$= \prod_{j=1}^{n}\prod_{i=1}^{k}\left[\frac{1}{\sqrt{2\pi}} e^{-(x_j-\mu_i)^2/2}\pi_i\right]^{z_j[i]}$$

Full data loglikelihood

$$\sum_{j=1}^{n}\sum_{i=1}^{k} z_j[i]\left[\frac{-(x_j-\mu_i)^2}{2} + \log \pi_i\right] + \text{const.}$$

$$q(z_1 \dots z_n) = q_1(z_1)\, q_2(z_2) \dots q_n(z_n)$$

$$\Lambda(q, \mu_1 \dots \mu_K, \pi) = \sum_{z_1 \dots z_n} q_1(z_1) \dots q_n(z_n) \sum_{j=1}^{n}\sum_{i=1}^{k} z_j[i]\left[-\frac{(x_j-\mu_i)^2}{2} + \log\pi_i\right]$$
$$- \sum_{z_1 \dots z_n} q_1(z_1) \dots q_n(z_n)\left[\log q_1(z_1) \dots q_n(z_n)\right].$$

Simplifying steps:

$$① \sum_{z_1 - z_n} q_1(z_1) \dots q_n(z_n)\left[\log q_1(z_1) \dots q_n(z_n)\right]$$

$$= \underbrace{\sum_{z_1 - z_n} q_1(z_1) \dots q_n(z_n) \log q_1(z_1)}_{||} + \dots$$

$$\sum_{z_1}\sum_{z_2 - z_n} q_1(z_1) q_2(z_2) \dots q_n(z_n) \log q_1(z_1)$$

$$= \sum_{z_1} q_1(z_1) \log q_1(z_1)\underbrace{\left[\sum_{z_2 \dots z_n} q_2(z_2) \dots q_n(z_n)\right]}_{=1}$$

so $① = \sum_{z_1} q_1(z_1)\log q_1(z_1) + \sum_{z_2} q_2(z_2) \log q_2(z_2) + \dots$

note here z_1 has k numbers.

$$\sum_{z_1 - z_n} q_1(z_1) \dots q_n(z_n) \sum_{j=1}^{n}\sum_{i=1}^{k} z_j[i]\left[\frac{-(x_j-\mu_i)^2}{2} + \log\pi_i\right]$$

$$= \sum_{z_1 - z_n} q_1(z_1) \dots q_n(z_n) \sum_{i=1}^{k} z_1[i]\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right].$$

$$+ \text{———} \sum_{i=1}^{k} z_2[i]\left[-\frac{(x_2-\mu_i)^2}{2} + \log\pi_i\right] + \dots$$

$$\to \sum_{z_1} q_1(z_1) \sum_{i=1}^{k} z_1[i]\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right], \text{ so}$$

$$\Lambda = \sum_{z_1} q_1(z_1) \sum_{i=1}^{k} z_1[i]\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right] - \sum_{z_1} q_1(z_1)\log q_1(z_1) +$$
$$+ \underbrace{\sum_{z_2} q_2(z_2) \sum_{i=1}^{k} z_2[i]\left[-\frac{(x_2-\mu_i)^2}{2} + \log\pi_i\right]}_{=h(z)} - \sum_{z_2} q_2(z_2) \log q_2(z_2) + \dots$$

by KL-divergence trick. $q_1(z_1) = \dfrac{e^{h(z_1)}}{\sum_r e^{h(r)}}$ ← q update.

probability of cluster j: $q_1(z_1 = [0 \dots \underset{j\text{th-component}}{1} \dots 0]) = \dfrac{\pi_j e^{-(x_1-\mu_j)^2/2}}{\sum_{\delta=1}^{k} \pi_\delta e^{-(x_1-\mu_\delta)^2/2}} = T_1[j]$

plugin: $$\sum_{z_1} q_1(z_1) \sum_{i=1}^{k} z_1[i]\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right]$$

$$= \sum_{i=1}^{k}\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right] \sum_{z_1} q_1(z_1) z_1[i].$$

$$= \left[-(x_1-\mu_1)^2/2 + \log\pi_1\right] \sum_{z_1} q_1(z_1) z_1[1] + \dots$$

$z_1[1]$ is 0 or 1. when $z_1 = [1,0,0,0]$ is 1

$= T_1[1]$.

$=$ prob $q_1(z_1 = [1, 0 \dots])$

so: $$\sum_{z_1} q_1(z_1) \sum_{i=1}^{k} z_1[i]\left[-\frac{(x_1-\mu_i)^2}{2} + \log\pi_i\right] = \sum_{i=1}^{k} T_1[i]\left[\frac{-(x_1-\mu_i)^2}{2} + \log\pi_i\right]$$

after q step, $\Lambda = \sum_{i=1}^{k} T_1[i]\left[-(x_1-\mu_i)^2/2 + \log\pi_i\right] +$ something with no μ or π

$+ \sum_{i=1}^{k} T_2[i]\left[-(x_2-\mu_i)^2/2 + \log\pi_i\right] +$ something $+ \dots$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k} T_j[i]\left(-(x_j-\mu_i)^2/2 + \log\pi_i\right) + \text{something}$$

$$\mu_1 = \frac{\sum_{j=1}^{n} T_j[1] x_j}{\sum_{j=1}^{n} T_j[1]}$$

μ_i update

$\mu_1 = \arg\max \sum_{j=1}^{n} T_j[1]\left[-(x_j-\mu_1)^2/2\right]$ | $\frac{\partial}{\partial \mu_1} = \sum_{j=1}^{n} T_j[1](x_j - \mu_1) = 0$

π update

$\pi = \arg\max \sum_{j=1}^{n}\sum_{i=1}^{k} T_j[i] \log\pi_i$ s.t. $\sum_{i=1}^{k}\pi_i = 1$. lagrange multiplier λ.

$$\sum_{j=1}^{n}\sum_{i=1}^{k} T_j[i]\log\pi_i + \lambda\left(\sum_{i=1}^{k}\pi_i - 1\right) \to \partial\pi_1 = \sum_{j=1}^{n} T_j[1]/\pi_1 + \lambda = 0$$

$$\Rightarrow \pi_1\lambda + \sum_{j=1}^{n} T_j[1] = 0.$$

$\partial\pi_2 = \dots \Rightarrow \pi_2\lambda + \sum_{j=1}^{n} T_j[2] = 0$, add them up using $\sum \pi_i = 1$

to get $\lambda + \underbrace{\sum_{j=1}^{n}\underbrace{\sum_{i=1}^{k} T_j[i]}_{=1}}_{=n} = 0$. so $\lambda + n = 0$. $\lambda = -n$.

so $\pi_1 = \sum_{j=1}^{n} T_j[1] / n$

multinomial$(n, \underset{\text{vector}}{p})$ $\underset{\text{vector of counts}}{P(k)} = \left(\dfrac{n!}{\prod_i k[i]!}\right)\prod_i p[i]^{k[i]}$, mean $= n p^{\text{vector}}$

$\Sigma = \text{Diag}(np) - npp^T$.

$E[(x[i]-\mu[i])(x[j]-\mu[j])] = -np[i]p[j]$.

$E[(x[i]-\mu[i])^2] = np[i](1-p[i])$.