Yihang Chen.

1. (a) suppose $g(u) = \log(1+e^{-u})$, we have.

$$g'(u) = -\frac{e^{-u}}{1+e^{-u}} \qquad g''(u) = \frac{e^{-u}(1+e^{-u}) + e^{-u}(-e^{-u})}{(1+e^{-u})^2}$$

$$= \frac{e^{-u}}{(1+e^{-u})^2} > 0$$

$\Rightarrow$ g is convex.

$$\Rightarrow \frac{\partial^2 f}{\partial \vec{x}^2} = \frac{\partial^2}{\partial \vec{x}} \sum_{i=1}^{n} g(b_i a_i^T \vec{x}) = \frac{\partial^2}{\partial \vec{x}} \sum_{i=1}^{n} g''(b_i a_i^T \vec{x}) b_i^2 a_i a_i^T$$

which is positive semidefinite $\Rightarrow$ f is convex

~~Since f is defined on $x \in \mathbb{R}^p$, and bounded below~~

The minimum does not always exist. eg. $f(x) = \log(1+e^{-x})$. the minimum is 0 and reached only by $x = +\infty$.

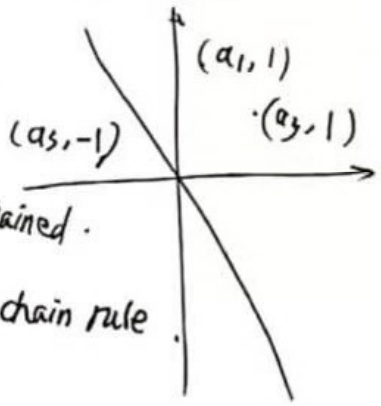
(b) ~~infima is~~ $f: \Omega \mapsto \mathbb{R}$. (i) a_{min} is the minimum $\Leftrightarrow$ $a_{min} \in \Omega$, $\forall a' \in \Omega$, $f(a') \geq f(a_{min})$

(ii) a_{inf} is infinima $\Leftrightarrow$ $\exists a^i \in \Omega$, $i \geq 1$. $a^i \to a_{inf}$, $\forall a' \in \Omega$, $f(a') \geq f(a_{inf})$.

The difference is that $a_{min} \in \Omega$, but a_{inf} might not in Ω.

$f(x) = \log(1+e^{-x})$. $x \to +\infty$, $f(x) \to 0$, does not attain infimum

(c). Say plane π is orthogonal to x^0, then $\{a_i\}$ is seperated by π, such that on one side it's 1, on the other side is -1.



~~Say $f(x)$~~ Say $g(\alpha) = f(\alpha x_0) = \sum_{i=1}^{n} \log(1+e^{-b_i \alpha a_i^T x_0})$

$\to 0$ as $\alpha \to +\infty$. so the minimum cannot be attained.

(d) $\nabla f_\mu(x) = \sum_{i=1}^{n} -b_i a_i \sigma(-b_i a_i^T x) + \mu x$ by the chain rule.

(e). $\nabla^2 f_\mu(x) = \sum_{i=1}^{n} b_i^2 \sigma''(-b_i a_i^T x) a_i a_i^T + \mu I$

$$= \sum_{i=1}^{n} \sigma(-b_i a_i^T x)(1 - \sigma(-b_i a_i^T x)) a_i a_i^T + \mu I.$$

since $\sigma''(t) = \frac{-e^{-t}(1+e^{-t})^2 + e^{-2t}(1+e^{-t})}{(1+e^{-t})^4} = \sigma(t)(1-\sigma(t))$

and $b_i^2 = 1$.

(f) Since $f_\mu - \frac{\mu}{2}\|x\|^2$ is convex, we have.

$$f_\mu(x) - \frac{\mu}{2}\|x\|^2 \geq f_\mu(y) - \frac{\mu}{2}\|y\|^2 + \nabla f_\mu(y)^T (y-x) + y^T(y-x)$$

$$= f_\mu(y) - \frac{\mu}{2}\|x\|^2 + \nabla f_\mu(y)^T (y-x) + \frac{\mu}{2}\|y-x\|^2 \Rightarrow f_\mu \text{ is } \mu\text{-strongly convex}$$

(g) (1) since $\text{rank}(a_i a_i^T) = 1 \Rightarrow a_i a_i^T$ has only one non-zero eigenvalue.

$\Rightarrow \lambda_{\max}(a_i a_i^T) = \text{tr}(a_i a_i^T) = \|a_i\|_2^2$.

(2) $\lambda_{\max}(\nabla^2 f_\mu(x)) = \lambda_{\max}\left(\sum_{i=1}^n a_i a_i^T \sigma(-b_i a_i^T x)(1-\sigma(-b_i a_i^T x))\right)$

$\le \lambda_{\max}\left(\sum_{i=1}^n a_i a_i^T + \lambda I\right)$ (by $0 \le \sigma(t)(1-\sigma(t)) \le 1$)

$\le \sum_{i=1}^n \lambda_{\max}(a_i a_i^T) + \mu = \sum_{i=1}^n \|a_i\|_2^2 + \mu$.

(3). Clearly ∇f_μ is continuously differentiable.

∇f_μ is L-smooth $\Leftrightarrow \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le L\|x-y\|_2$.

or by Taylor expansion, $\nabla f_\mu(x) = \nabla f_\mu(y) + \nabla^2 f_\mu(\alpha x + (1-\alpha)y)(x-y)$

$\therefore \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_2 \le \|\nabla^2 f_\mu(\alpha x + (1-\alpha)y)\|_2 \|x-y\|_2$

$\le \lambda_{\max}(\nabla^2 f_\mu)\|x-y\|_2 = \left(\|A\|_F^2 + \mu\right)\|x-y\|_2$

2.2. (a) according to 1.(g)(2), f_i is $(\|a_i\|^2+\mu)$-Lip continuous, and hence $L_{\max} = \max_i L(f_i)$ -Lip continuous.

Then $\mathbb{E}_{i\sim \text{Uniform}([1,n])} \nabla f_i(x) = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x) = \nabla f(x)$.

$\Rightarrow \nabla f_i$ is an unbaised estimate of $\nabla f(x)$

2.3. (b) Take the subgradient: $\lambda \nabla g(y) + y - z = 0 \Rightarrow z = y + \lambda \nabla g(y)$

$\forall$ coordinate i, $z_i = y_i + \lambda \nabla(\|y\|_1)_i$

since $\nabla|x| = \begin{cases} 1 & x>0 \\ [-1,1] & x=0 \\ -1 & x<0 \end{cases}$ we have

if $z_i > \lambda$. y_i must >0. $\Rightarrow y_i = z_i - \lambda$.

if $z_i < -\lambda$ y_i must <0 $\Rightarrow y_i = z_i + \lambda$.

~~if $|z_i| \le \lambda$~~ if $0 \le z_i \le \lambda$. and if $y_i \not> 0$. $\Rightarrow y_i + \lambda > \lambda \ge z_i$. contradiction

similarly $y_i < 0$ does not hold $\Rightarrow y_i = 0$.

similarly if $-\lambda \le z_i \le 0 \Rightarrow y_i = 0$.

in sum, $\text{prox}_{\lambda g}(z) = y = \text{sign}(z) \circ \max(|z| - \lambda, 0)$

The solution before Part 2.4 is hand-written in anther PDF file.

2.1

I implemet all the credited method in algorithm.py

2.2/2.3

It is contained in the written part.

2.4

Go to folder ./question2.

a

The observed convergence rates is not entirely consistent with the theoretical ones. The theoretical ones provide an upper bound. I adopt a tighter bound by noticing $\sigma(t)(1-\sigma(t)) \leq \frac{1}{4}$, since we could have

$$\|\nabla f_\mu(x)\|_2 \leq \frac{1}{4}\|A^\top A\|_2 + \mu$$

The results are shown in Line 33 in question_2_4.py and *_tight.pdf files. We observe that it is a tighter estimate than before.

b

The observed convergence rates of GD and GDstr is linear. Since our objective is strongly convex, setting $\alpha = \frac{1}{L}$ and $\alpha = \frac{2}{L+\mu}$ will result in linear convergence. The regression result confirms such idea.

d

The empirical speed of convergence is faster than the theoretical one. Since the theoretical result is an upper bound. The bash output for ./question_2_4.py is

GD results: a_GD=-0.000180, b_GD=1.196370 GD theoretical rate: a_GD=-0.000053, b_GD=1.282497 GD rate diff =-0.00012623199407315507 GD_str results: a_GD=-0.000289, b_GD=1.119041 GD_str theoretical rate: a_GD=-0.000107, b_GD=1.282084436227831 GD_str rate diff =-0.000182

and the resulting figures are stored in ./figs/fig_ex2_4_convergence_rate,pdf and ./figs/fig_ex2_4_convergence_rate_full.pdf. If we use a tighter upper bound as discussed in 2.4.a, the resulting figures are stored in ./figs/fig_ex2_4_convergence_rate_tight,pdf and ./figs/fig_ex2_4_convergence_rate_full_tighter.pdf

All the method is tested in ./log_reg.py, the classification error for ./log_reg.py is reported below:

GD : 0.1386861313868613

GDstr : 0.0948905109489051

AGD : 0.058394160583941604

AGDstr : 0.058394160583941604

AGDR : 0.072992700729927

AdaGrad : 0.058394160583941604

SGD : 0.072992700729927

SAG : 0.051094890510948905

SVR : 0.058394160583941604

SubG : 0.12408759124087591

L1-prox

ISTA : 0.1386861313868613

FISTA : 0.145985401459854

FISTAR : 0.145985401459854

PROXSG : 0.145985401459854

L2-prox

ISTA : 0.11678832116788321

FISTA : 0.06569343065693431

FISTAR : 0.058394160583941604

PROXSG : 0.12408759124087591

we can see that FISTA and FISTAR, AdaGrad, Stochastic method (SGD, SAG, SVR) would be better choice.

3.3.1

(a)

$$\nabla(f_{\ell_1} + g_{\ell_1})(\alpha) = \mathbf{W}\mathbf{P}_\Omega^\top(\mathbf{P}_\Omega\mathbf{W}^\top\alpha - \mathbf{b}) + \lambda_{\ell_1}\text{sign}(\alpha)$$

where

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$$

and applied element-wisely. Then

$$\nabla(f_{\mathbf{TV}} + g_{\mathbf{TV}})(x) = \mathbf{P}_\Omega^\top(\mathbf{P}_\Omega x - b) + \lambda_{\mathbf{TV}}\nabla g_{\mathbf{TV}}(x)$$

- For the isotropic case. Define

$$D_1 = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ 0 & & \cdots & & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}$$

(1) We start with the differencing matrix by column. It is just the differencing matrix of a TV term of dimension m, namely, D_1, i.e., $\nabla x^1_{:,j} = D_1 \cdot x_{:,j}$.

(2) We can then proceed by column, i.e., $\nabla x_{i,:} = x_{i,:} \cdot D_1^\top$.

(3) If we stacking the matrix into a vector, define $\text{vec}(X)_{i+(j-1)m} = X_{i,j}$, we could have $\text{vec}(\nabla x^1) = (I \otimes D_1) \cdot \text{vec}(x)$, where I is a $m \times m$ identity matrix.

(4) Since two consecutive entries of the same row are separated by m positions after vectorization, we have $\text{vec}(\nabla x^2) = (D_1 \otimes I) \cdot \text{vec}(x)$.

(5) In total, we can write $\|x\|_{\mathbf{TV},\ell_1} = \|(I \otimes D_1) \cdot \text{vec}(x)\|_1 + \|(D_1 \otimes I) \cdot \text{vec}(x)\|_1$ hence the gradient is
$\nabla \|x\|_{\mathbf{TV},\ell_1} = (I \otimes D_1^\top)\text{sign}((I \otimes D_1) \cdot \text{vec}(x)) + (D_1^\top \otimes I)\text{sign}((D_1 \otimes I) \cdot \text{vec}(x))$
by the chain rule.

- For the anisotropic case, the matrix form does not simplify the calculation. Denote $x_g = \nabla \|x\|_{\mathbf{TV},\ell_2}$. Notice $x_{i,j}$ only appears in the term $\nabla x^1_{i,j}$, $\nabla x^2_{i,j}$, $\nabla x^1_{i-1,j}$, $\nabla x^2_{i,j-1}$. We have

$$(\nabla \|x\|_{\mathbf{TV},\ell_2})_{i,j} = \frac{\partial(\|\nabla x^1_{i,j}\|_2 + \|\nabla x^2_{i,j}\|_2 + \|\nabla x^1_{i-1,j}\|_2 + \|\nabla x^2_{i,j-1}\|_2)}{\partial x_{i,j}} = -\frac{\nabla x^1_{i,j} + \nabla x^2_{i,j}}{\|\nabla x_{i,j}\|_2} + \frac{\nabla x^1_{i-1,j}}{\|\nabla x_{i-1,j}\|_2}$$

which is the gradient.

(b)

By the definition of $\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$. And by SVD decomposition, $A = U\Sigma V^\top$ we have $A^\top A = V\Sigma^2 V^\top$. Hence, the operator norm of $A^\top A$ (equals the largest singular value) is the square of the operator norm of A.

Since

$$\nabla f_{\ell_1}(\alpha) = \mathbf{W}\mathbf{P}_\Omega^\top(\mathbf{P}_\Omega \mathbf{W}^\top \alpha - \mathbf{b})$$

Hence, the Lipschitz constant would be

$$\|\mathbf{W}\mathbf{P}_\Omega^\top \mathbf{P}_\Omega \mathbf{W}^\top\|_2 = \|\mathbf{P}_\Omega \mathbf{W}^\top\|_2^2 = \|\mathbf{P}_\Omega\|_2^4 = 1$$

since $\mathbf{W}$ is unitary matrix, and $\mathbf{P}_\Omega \mathbf{P}_\Omega^\top$ is a diagnoal matrix which only has 1 as its nonzero element on its diagonal. Similarly, the Lipschitz constant for $\nabla f_{\mathbf{TV}}$ is

$$\|\mathbf{P}_\Omega^\top \mathbf{P}_\Omega\|_2 = \|\mathbf{P}_\Omega\|_2^2 = 1$$

3.3.2 Go to folder ./question3. The reconstructed images are plotted in ./results/log_lambda_{lambda}.png, when setting lambda = $10^{\{-3,-2.5,-2,\cdots,0.5,1\}}$, and the PSNR-lambda relation is plotted in ./results/PSNR-lambda.png. We can see that setting lambda around 0.01-0.1 would be the best choice. Setting λ too large will make ℓ_1 penalized method output zero.

In general, from the recovered image, we find that ℓ_1 norm tends to make the reconstructed image dimmer than the original ones, and it cannot make clear distinctions between different patch of color, but it can recover some details of the original image. While the reconstruction image by TV norm will lose detailed information, but has almost the same brightness, and can make clear distinction between dark and light colors.