Optimization for Data Science

2018

Definition 0.1: $\sigma_{\min}$ **and** $\sigma_{\max}$

Let $A \in \mathbb{R}^{d\times d}$ be a matrix and let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A defined by:

$$\sigma_{\min}(A) \overset{def}{=} \min_{x\in\mathbb{R}^d, x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} \quad \text{and} \quad \sigma_{\max}(A) \overset{def}{=} \max_{x\in\mathbb{R}^d, x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} \tag{0.0.1}$$

Proposition 0.1: $\sigma_{\max}$ **of a symetric positive semi-definite matrix**

If A is a symmetric positive semi-definite matrix:

$$\sigma_{\max}(A) = \max_{x\in\mathbb{R}^d, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^d, x\neq 0} \frac{\langle Ax, x\rangle}{\|x\|_2^2} \tag{0.0.2}$$

Therefore :

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \sigma_{\max}(A) \quad \forall x \in \mathbb{R}^d \tag{0.0.3}$$

and

$$\frac{\langle Ax, x\rangle}{\|x\|_2^2} \leq \sigma_{\max}(A) \quad \forall x \in \mathbb{R}^d \tag{0.0.4}$$

Warm-up : Proving convergence of the Gradient Descent Method on the Ridge Regression Problem.

$$f(w) \overset{def}{=} \frac{1}{2n}\|X^T w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \tag{0.0.5}$$

We will now solve the following ridge regression problem :

$$w^\star = \arg\min_{w\in\mathbb{R}^d} (f(w)) \tag{0.0.6}$$

using gradient descent :

$$w^{t+1} = w^t - \alpha \nabla f(w^t) \tag{0.0.7}$$

where:

$$\alpha = \frac{1}{\sigma_{\max}(A)} \tag{0.0.8}$$

with:

$$A \overset{def}{=} \frac{1}{n} X X^T + \lambda I \tag{0.0.9}$$

Exercise 0.1. Show that $\nabla f(x)$ is given by

$$\nabla f(x) = Aw - b = A(w - w^\star)$$

where $w^\star$ is the solution of (**??**) and

$$b \overset{def}{=} \frac{1}{n} X y \tag{0.0.10}$$

Proof. As a reminder,

$$\nabla \|Ax + b\|^2 = 2A^T(Ax + b) \tag{0.0.11}$$

$$\begin{aligned}
\nabla f(w) &\overset{(??)}{=} \nabla(\frac{1}{2n}\|X^T w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2) \\
&\overset{(??)}{=} \frac{1}{n}X(X^T w_y) + \lambda w \\
&= (\frac{1}{n}XX^T + \lambda I)w - \frac{1}{n}Xy \\
&= Aw - b
\end{aligned} \tag{0.0.12}$$

$w^\star$ is a solution of (**??**) implies :

$$\nabla f(w^\star) = 0 \overset{(??)}{\implies} b = Aw^\star \tag{0.0.13}$$

$$(??) + (??) \implies \nabla f(w) = Aw - b = A(w - w^\star) \tag{0.0.14}$$

□

Exercise 0.2. Show that A as defined in (**??**) is positive semi-definite, that is

$$\langle Aw, w\rangle \geq 0, \forall w \in \mathbb{R}^d \tag{0.0.15}$$

and that

$$\sigma_{\max}(I - \alpha A) = 1 - \alpha\sigma_{\min}(A) = 1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)} \tag{0.0.16}$$

Proof.

$$\begin{aligned}
\langle Aw, w\rangle &= w^T Aw \\
&= w^T(\frac{1}{n}XX^T + \lambda I)w \\
&= \frac{1}{n}w^T XX^T w + w^T w \\
&= \frac{1}{n}\|X^T w\| + \|w\| \geq 0
\end{aligned}$$

$$A \succeq 0 \text{ and symmetric } \overset{(??)}{\implies} \langle Aw, w\rangle \leq \sigma_{\max}(A)\|w\|^2$$

$$\begin{aligned}
\langle (I - \alpha A)w, w\rangle &= \|w\|^2 - \alpha\langle Aw, w\rangle \\
&\geq \|w\|^2 - \alpha(\sigma_{\max}(A)\|w\|^2) \\
&\geq 0 \implies (I - \alpha A) \succeq 0
\end{aligned}$$

$$\begin{aligned}
(I - \alpha A) \succeq 0 \text{ and symmetric } \overset{(??)}{\implies} \sigma_{\max}(I - \alpha A) &= \max_w \frac{\langle (I - \alpha A)w, w\rangle}{\|w\|^2} \\
&= \max_w \frac{\|w\|^2 - \alpha\langle Aw, w\rangle}{\|w\|^2} \\
&= 1 - \alpha \min_w \frac{\langle Aw, w\rangle}{\|w\|^2} \\
&= 1 - \alpha\sigma_{\min}(A) \\
&= 1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}
\end{aligned}$$

□

Exercise 0.3. Show that the iterates (**??**) converge to $w^\star$ according to

$$\|w^{t+1} - w^\star\|_2 \leq (1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)})\|(w^t - w^\star)\|_2$$

Proof. Using (**??**)

$$
\begin{aligned}
w^{t+1} &= w^t - \alpha A(w^t - w^\star) \\
w^{t+1} - w^\star &= (I - \alpha A)(w^t - w^\star)
\end{aligned}
$$

Taking norms

$$
\|w^{t+1} - w^\star\| = \|(I - \alpha A)(w^t - w^\star)\| \tag{0.0.17}
$$

$$
\|(I - \alpha A)x\|_2 \overset{(??)}{\leq} \sigma_{\max}(I - \alpha A)\|x\|_2
$$

taking $x = w^t - w^\star$

$$
\|(I - \alpha A)(w^t - w^\star)\|_2 \overset{(??)}{\leq} \sigma_{\max}(I - \alpha A)\|(w^t - w^\star)\|_2
$$

With (**??**)

$$
\|w^{t+1} - w^\star\| \leq \sigma_{\max}(I - \alpha A)\|(w^t - w^\star)\|_2
$$

□

Exercise 0.4. Let

$$
\kappa(A) \overset{def}{=} \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}
$$

which is known as the condition number of A. What happens to κ as $\lambda \to \infty$ and $\lambda \to 0$, respectively? What does this imply about the speed at which gradient descent converges to the solution?

Proof. Note that

$$
\sigma_{\max}(\frac{1}{n}XX^T + \lambda I) = \frac{1}{n}\sigma_{\max}^2(X) + \lambda
$$

Therefore we have $\kappa =$

$$
\begin{aligned}
\kappa &= \frac{\frac{1}{n}\sigma_{\min}^2(X) + \lambda}{\frac{1}{n}\sigma_{\max}^2(X) + \lambda} \\
&\underset{\lambda \to \infty}{\rightarrow} 1 \\
&\underset{\lambda \to 0}{\rightarrow} \kappa(X)^2
\end{aligned}
$$

□

1 Properties and examples of convexity and smoothness.

Notation : For every $x, y \in \mathbb{R}^d$ let $\langle x, y, \rangle \overset{def}{=} x^T y$ and $\|x\|_2 = \sqrt{\langle x, x \rangle}$. Let $\sigma_{min}(A)$ and $\sigma_{max}(A)$ be the smallest and largest singular values of A defined by:

$$\sigma_{min}(A) \overset{def}{=} \min_{x \in \mathbb{R}^d} \frac{\|Ax\|_2}{\|x\|_2} \text{ and } \sigma_{\max}(A) \overset{def}{=} \max_{x \in \mathbb{R}^d} \frac{\|Ax\|_2}{\|x\|_2} \tag{1.0.1}$$

Thus clearly :

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \sigma_{\max}(A)^2, \forall x \in \mathbb{R}^d. \tag{1.0.2}$$

Let $\|A\|_F^2 \overset{def}{=} Tr(A^T A)$ denote the Frobenius norm of A.
For every symmetric matrix G the $L2$ induced matrix norm can be equivalently defined by :

$$\|G\|_2 = \sigma_{max}(G) = \sup_{x \in \mathbb{R}^d, x \neq 0} \frac{|\langle Gx, x \rangle|}{\|x\|_2^2} = \max_{x \in \mathbb{R}^d, x \neq 0} \frac{\|Gx\|_2}{\|x\|_2} \tag{1.0.3}$$

1.1 Convexity

1.1.1 Lecture

Definition 1.1: Convexity

We say that a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if:

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y), \forall x, y \in \mathbb{R}^d, \lambda \in [0, 1]. \tag{1.1.1}$$

Proposition 1.1: Convexity : first derivate

A differential function $f : dom(f) \subset \mathbb{R}^n \to \mathbb{R}$ is convex iff :

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle \tag{1.1.2}$$

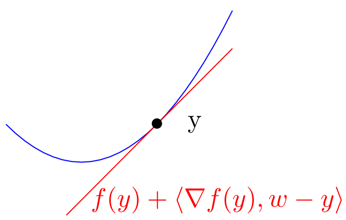

Figure 1: Convexity : first derivate

Proof.

$$\begin{aligned}
(1.1.1) &\Leftrightarrow \lambda(f(x) - f(y)) \geq f(y + \lambda(x - y)) - f(y) \\
&\Leftrightarrow f(x) \geq f(y) + \frac{f(y + \lambda(x - y)) - f(y)}{\lambda} \\
&\Leftrightarrow (1.1.2) \text{ with } \lambda \to 0
\end{aligned}$$

□

Proposition 1.2: Convexity : second derivate

A differential function $f : dom(f) \subset \mathbb{R}^n \to \mathbb{R}$ is convex iff :

$$v^T \nabla^2 f(w) v \geq 0, \Leftrightarrow \nabla^2 f(w) \succeq 0, \forall w, v \in \mathbb{R}^n \tag{1.1.3}$$

Proof. Using Taylor's expansion:

$$f(a+h) = f(a) + \nabla f(a)^T h + \frac{1}{2} h^T \nabla^2(a) h + o(\|h\|^2) \tag{1.1.4}$$

Subsituting $w = a + h$ and $y = a$:

$$f(w) = f(y) + \nabla f(y)^T (w-y) + \frac{1}{2}(w-y)^T \nabla^2(y)(w-y) + o(\|w-y\|^2)$$

And using (1.1.2). □

Definition 1.2: Strong convexity

We say that f is μ-strongly convex if :

$$f(w) \geq f(y) + \langle \nabla f(y), w-y \rangle + \frac{\mu}{2}\|w-y\|^2, \forall w, y \in \mathbb{R}^d. \tag{1.1.5}$$

or

$$v^T \nabla^2 f(x) v \geq \mu \|v\|_2^2, \forall x, v \in \mathbb{R}^d \tag{1.1.6}$$

Proposition 1.3: Polyak-Lojasiewicz inequality

If $f : \mathbb{R}^n \to \mathbb{R} \cup \infty$ is μ-strongly convex then

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^\star)), \forall x \in \mathbb{R}^n \tag{1.1.7}$$

Proof. With $w = y - \frac{1}{\mu}\nabla f(y)$

$$\begin{aligned}
f(y - \frac{1}{\mu}\nabla f(y)) - f(y) &\overset{(1.1.5)}{\geq} \langle \nabla f(y), -\frac{1}{\mu}\nabla f(y) \rangle + \frac{\mu}{2}\| - \frac{1}{\mu}\nabla f(y)\|^2 \\
&\geq -\frac{1}{\mu}\|\nabla f(y)\|^2 + \frac{1}{2\mu}\|\nabla f(y)\|^2 \\
&\geq -\frac{1}{2\mu}\|\nabla f(y)\|^2
\end{aligned}$$

Therefore

$$\begin{aligned}
\|\nabla f(y)\|^2 &\geq -2\mu(f(y - \frac{1}{\mu}\nabla f(y)) - f(y)) \\
&\geq 2\mu(f(y) - f(y^\star))
\end{aligned}$$

□

Proposition 1.4: Convexity Properties

1. $x \mapsto \|x\|$ is a convex function.
2. If f convex, $g : x \in \mathbb{R}^d \mapsto f(Ax - b)$ is convex.
3. If $f_i : \mathbb{R}^d \to \mathbb{R}$ convex for $i = 1, \ldots, m$, $\sum_{i=1}^m f_i$ is convex.
4. Let $A \in \mathbb{R}^{m \times d}$ have full column rank. $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ is $\sigma_{\min}(A)$-strongly convex.

1.1.2 Exercises

Definition 1.3: Norm

We say that $\|.\| \to \mathbb{R}_+$ is a norm over $\mathbb{R}^d$ if it satisfies the following three properties:
1. Point separating: $\|x\| = 0 \Leftrightarrow x = 0, \forall x \in \mathbb{R}^d$
2. Subadditive: $\|x + y\| \le \|x\| + \|y\|, \forall x, y \in \mathbb{R}^d$
3. Homogeneous: $\|ax\| = |a|\|x\|, \forall x \in \mathbb{R}^d, a \in \mathbb{R}$

Exercise 1.1. Prove that $x \mapsto \|x\|$ is a convex function.

Proof. Let $\lambda \in [0,1], x, y \in \mathbb{R}^d$.

$$\begin{aligned}\|\lambda x + (1-\lambda)y\| &\overset{item2}{\le} \|\lambda x\| + \|(1-\lambda)y\| \\ &\overset{item3}{\le} \lambda\|x\| + (1-\lambda)\|y\|\end{aligned}$$

□

Exercise 1.2. For every convex function $f : y \in \mathbb{R}^m \mapsto f(y)$, prove that $g : x \in \mathbb{R}^d \mapsto f(Ax - b)$ is a convex function, where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$.

Proof. Let $\lambda \in [0,1], x, y \in \mathbb{R}^d]$.

$$\begin{aligned}g(\lambda x + (1-\lambda)y) &= f(A(\lambda x + (1-\lambda))y - b) \\ &= f(\lambda(Ax - b) + (1-\lambda)(Ay - b)) \qquad [\text{using } b = \lambda b + (1-\lambda)b] \\ &\overset{f \text{ is convex}}{\le} \lambda f(Ax - b) + (1-\lambda)f(Ay - b)\end{aligned}$$

□

Exercise 1.3. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be convex for $i = 1, \dots, m$. Prove that $\sum_{i=1}^m f_i$ is convex.

Proof. Immediate through either definition. □

Exercise 1.4. For given scalars $y_i \in \mathbb{R}$ and vectors $a_i \in \mathbb{R}^d$ for $i = 1, \dots, m$ prove that the logistic regression function $f(x) = \sum_{i=1}^m \ln(1 + e^{-y_i\langle x, a_i\rangle})$ is convex.

Proof. From Exercice 3 we only need prove that $f(x) = \ln(1 + e^{-y\langle x, a_i\rangle})$ is convex for a given $y \in \mathbb{R}$ and $w \in \mathbb{R}^d$.
From Exercice 2 we only need prove that $\phi(\alpha) = \ln(1 + e^\alpha)$ is convex, since $x \mapsto -y\langle x, w\rangle$ is a linear function.
We have:

$$\phi'(\alpha) = \frac{e^\alpha}{1 + e^\alpha}$$

$$\text{and differentiating again} \quad \phi''(\alpha) = \frac{e^\alpha(1 + e^\alpha) - e^{2\alpha}}{(1 + e^\alpha)^2} = \frac{e^\alpha}{(1 + e^\alpha)^2} \ge 0$$

Using the definition (1.1.3) prove that ϕ is convex. □

Exercise 1.5. Let $A \in \mathbb{R}^{m \times d}$ have full column rank. Prove that $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ is $\sigma_{\min}(A)$-strongly convex.

Proof.

$$\nabla f(x) = \frac{1}{2}2A^T(Ax - b) = A^T(Ax - b)$$
$$\nabla^2 f(x) = A^T A$$

And

$$v^T \nabla^2 f(x) v = v^T A^T A v = \|Av\|_2^2 \overset{(1.0.1)}{\ge} \sigma_{\min}(A)\|v\|_2^2$$

□

1.2 Smoothness

1.2.1 Lecture

Definition 1.4: Smoothness

$f : \mathbb{R}^n \to \cup\{\infty\}$ is L-smooth if:

$$f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2}\|w - y\|^2, \forall w, y \in \mathbb{R}^n \tag{1.2.1}$$

Proposition 1.5: Smoothness Equivalence

A twice differentiable $f : \mathbb{R}^n \to \cup\{\infty\}$ is L-smooth if either:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n \tag{1.2.2}$$

$$d^T \nabla^2 f(x) d \leq L\|d\|^2, \forall x, d \in \mathbb{R}^n \tag{1.2.3}$$

Proof. let prove (1.2.3) $\implies$ (1.2.2):

$$\begin{aligned} d^T \nabla^2 f(x) d \leq L\|d\|^2 &\Leftrightarrow \frac{d^T \nabla^2 f(x) d}{\|d\|^2} \leq L \\ &\Leftrightarrow \frac{\langle \nabla^2 f(x) d, d \rangle}{\|d\|^2} \leq L \\ &\overset{\nabla^2 \text{is symmetric} + (1.0.3)}{\implies} \sigma_{\max}(\langle \nabla^2 f(x)) \leq L \end{aligned} \tag{1.2.4}$$

Using first Taylor's expansion on ∇f,

$$\begin{aligned} \|\nabla(f(x+v)) - \nabla f(x)\| &= \|\int_0^1 \nabla^2 f(x + \alpha v) v \partial\alpha\| \\ &\overset{\|\int\| \leq \int \|\cdot\|}{\leq} \int_0^1 \|\nabla^2 f(x + \alpha v) v\| \partial\alpha \\ &\overset{(1.0.1)}{\leq} \int_0^1 \sigma_{\max}(\nabla^2 f(x + \alpha v))\|v\| \partial\alpha \\ &\overset{(1.2.4)}{\leq} \int_0^1 L\|v\| \partial\alpha \\ &\leq L\|v\| \end{aligned}$$

$\square$

Proof. let prove (1.2.3) $\implies$ (1.2.1):
Second Taylor's expansion :

$$f(x) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2}\int_0^1 (x - y)^T \nabla^2 f(x + \alpha(y - x))(x - y) \partial\alpha \tag{1.2.5}$$

$$(x - y)^T \nabla^2 f(x + \alpha(y - x))(x - y) \overset{(1.2.3)}{\leq} L\|x - y\|^2 \tag{1.2.6}$$

$$(1.2.5) + (1.2.6) \implies (1.2.1)$$

$\square$

Proposition 1.6: Smoothness Property

If $f : \mathbb{R}^n \to \mathbb{R} \cup \infty$ is L-smooth then :

$$f(w - \frac{1}{L}\nabla f(w)) - f(w) \leq -\frac{1}{2L}\|\nabla f(w)\|_2^2, \forall w \in \mathbb{R}^n \tag{1.2.7}$$

Because $w^\star \leq w$:

$$f(w^\star) - f(w) \leq -\frac{1}{2L}\|\nabla f(w)\|_2^2, \forall w \in \mathbb{R}^n \tag{1.2.8}$$

Proof. Substituting $y = w - \frac{1}{L}\nabla f(w)$:

$$\begin{aligned}
f(w - \frac{1}{L}\nabla f(w)) - f(w) &= f(y) - f(w) \\
&\overset{(1.2.1)}{\leq} \langle \nabla f(w), y - w\rangle + \frac{L}{2}\|y - w\|^2 \\
&\leq \langle \nabla f(w), -\frac{1}{L}\nabla f(w)\rangle + \frac{L}{2}\| - \frac{1}{L}\nabla f(w)\|^2 \\
&\leq -\frac{1}{L}\|\nabla f(w)\|^2 + \frac{1}{2L}\|\nabla f(w)\|^2
\end{aligned}$$

□

Proposition 1.7: Smoothness Properties

1. If $f : \mathbb{R} \to \mathbb{R}^d$ twice differentiable and L-smooth, $\sigma_{\max}(\nabla^2 f(x)) = \|\nabla^2 f(x)\|_2 \leq L$
2. If $f : \mathbb{R} \to \mathbb{R}^d$ twice differentiable and L-smooth,$g : x \in R^d \mapsto f(Ax - b)$ is $L\|A\|^2$-smooth.
3. If $f_i : \mathbb{R}^d \to \mathbb{R}$ twice differentiable and L_i-smooth for $i = 1, \dots, m$, $g = \frac{1}{n}\sum f_i$ is $(\sum \frac{L_i}{n})$-smooth.
4. $f : x \mapsto \frac{1}{2}\|Ax - b\|_2^2$ is $\sigma_{\max}^2(A)$-smooth.

1.2.2 Exercises

Exercise 2.2. Let $f : \mathbb{R} \to \mathbb{R}^d$ be twice differentiable and L-smooth. Show that:

$$\sigma_{\max}(\nabla^2 f(x)) = \|\nabla^2 f(x)\|_2 \leq L \tag{1.2.9}$$

Proof.

$$\nabla^2 \text{symetric} + (1.0.3) \implies$$

$$\begin{aligned}
\|\nabla^2 f(x)\|_2 = \sigma_{max}(\nabla^2 f(x)) &= \sup_{v \in \mathbb{R}^d, v \neq 0} \frac{|\langle \nabla^2 f(x) v, v|}{\|v\|_2^2} \\
&\overset{(1.2.3)}{=} \sup_{v \in \mathbb{R}^d, v \neq 0} \frac{L\|v\|_2^2}{\|v\|_2^2} = L
\end{aligned}$$

□

Exercise 2.3. For every twice differentiable L-smooth function $f : y \in \mathbb{R}^m \mapsto f(y)$, prove that $g : x \in R^d \mapsto f(Ax - b)$ is a smooth function, where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. Find the smoothness constant of g.

Proof.

$$\nabla g(x) = A^T \nabla f(Ax - b)$$

Therefore:

$$\begin{aligned}
\|\nabla g(x) - \nabla g(y)\| &= \|A^T(\nabla f(Ax - b) - \nabla f(Ay - b))\| \\
&\leq \|A^T\|\|\nabla f(Ax - b) - \nabla f(Ay - b)\| \\
&\overset{(1.2.2)}{\leq} L\|A^T\|\|A(x - y)\| \\
&\leq L\|A^T\|\|A\|\|x - y\|
\end{aligned}$$

g is $L\|A\|^2$-smooth. □

Exercise 2.4. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable and L_i-smooth for $i = 1, \dots, m$. Prove that $g = \frac{1}{n}\sum f_i$ is $(\sum \frac{L_i}{n})$-smooth.

Proof.

$$\begin{aligned}
\|\nabla^2 g(x)\| &= \|\nabla^2 \frac{1}{n}\sum f_i(x)\| \\
&= \|\frac{1}{n}\sum \nabla^2 f_i(x)\| \\
&\overset{\text{subadditivity of}\|\|}{\leq} \frac{1}{n}\sum \|\nabla^2 f_i(x)\| \\
&\overset{(1.2.9)}{\leq} \frac{1}{n}\sum L_i
\end{aligned}$$

□

Exercise 2.5. For given scalars $y_i \in \mathbb{R}$ and vectors $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$, prove that the logistic regression function $f(x) = \sum_{i=1}^m \ln(1 + e^{-y_i\langle x, a_i\rangle})$ is smooth.

Proof. $\phi(\alpha) = \ln(1 + e^\alpha)$ is twice differentiable, with :

$$\phi'(\alpha) = \frac{e^\alpha}{1+e^\alpha}$$

$$\phi''(\alpha) = \frac{e^\alpha(1+e^\alpha) - e^{2\alpha}}{(1+e^\alpha)^2} = \frac{e^\alpha}{(1+e^\alpha)^2} \leq 1$$

$$(1.2.9) \implies \phi \text{ is at least 1-smooth}$$

$x \mapsto -y\langle x, a\rangle$ is a linear function, also :

$$\text{Exercice 2.3.} \implies x \mapsto \ln(1 + e^{-y\langle x,a\rangle}) \text{ is at least } y^2\|w\|^2\text{-smooth}$$

Finally

$$\text{Exercice 2.4.} \implies g \text{ is at least } (\frac{1}{m}\sum y_i^2\|a_i\|^2)\text{-smooth}$$

□

Exercise 2.6. Let $A \in \mathbb{R}^{m\times d}$ be any matrix. Prove that $f : x \mapsto \frac{1}{2}\|Ax - b\|_2^2$ is $\sigma^2_{\max}(A)$-smooth.

Proof.

$$\nabla^2 f(x) = \nabla(A^T(Ax - b)) = A^T A$$

Consequently

$$v^T\nabla^2 f(x)v = v^T A^T A v = \|Av\|^2 \leq \sigma^2_{\max}(A)\|v\|^2$$

□

Exercise 2.7. Let $M > 0$ be a positive constant. Let $f(x) = \frac{1}{n}\sum_{i=1}^n \phi_i(a_i^T x)$ is a scalar function such that $\phi_i''(t) \leq M, \forall t \in \mathbb{R}$. Prove that f is $M\sigma^2_{\max}(A)$-smooth.

Proof. Using $\nabla a^T x = a$, we have $\nabla g(a^T x) = ag'(a^T x)$

$$\begin{aligned}
\nabla f(x) &= \frac{1}{n}\sum_{i=1}^n \nabla\phi_i(a_i^T x) \\
&= \frac{1}{n}\sum_{i=1}^n a_i\phi_i'(a_i^T x)
\end{aligned}$$

so

$$\begin{aligned}
\nabla^2 f(x) &= \frac{1}{n}\sum_{i=1}^n \nabla(\phi_i'(a_i^T x)a_i) \\
&= \frac{1}{n}\sum_{i=1}^n a_i^T\phi_i''(a_i^T x)a_i \\
&= A^T\Phi(x)A, \text{ where } \Phi(x) = diag(\phi_1''(a_1^T x), \ldots, \phi_n''(a_n^T x))
\end{aligned}$$

Consequently :

$$\|\nabla^2 f(x)\| = \|A^T\Phi(x)A\| \leq \|A\|^2\|\Phi(x)\| \leq M\|A\|^2$$

□

2 Gradient Descent.

Consider the problem :

$$w^\star = \arg\min_{w \in \mathbb{R}^d} (f(w)) \tag{2.0.1}$$

and the following gradient method:

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

2.1 Gradient Descent if f μ-strong convex and L-smooth.

Theorem 2.1: Convergence GD 1

Let f be μ-convex and L-smooth.

$$\|w^T - w^\star\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|w^1 - w^\star\|_2^2 \tag{2.1.1}$$

where $w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$ for $t = 1, \ldots, T$.

$$\implies \text{ for } \frac{\|w^T - w^\star\|_2^2}{\|w^1 - w^\star\|_2^2} \leq \epsilon \text{ we need } T \geq \frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

Proof.

$$\begin{aligned}
\|w^{t+1} - w^\star\|_2^2 &= \|w^t - \frac{1}{L}\nabla f(w^t) - w^\star\|^2 \\
&= \|(w^t - w^\star) - \frac{1}{L}\nabla f(w^t)\|^2 \\
&= \|(w^t - w^\star)\|^2 - 2\langle \frac{1}{L}\nabla f(w^t), w^t - w^\star\rangle + \|\frac{1}{L}\nabla f(w^t)\|^2 \\
&= \|(w^t - w^\star)\|^2 - \frac{2}{L}\langle \nabla f(w^t), w^t - w^\star\rangle + \frac{1}{L^2}\|\nabla f(w^t)\|^2 \\
&\overset{(1.1.5)}{\leq} \|(w^t - w^\star)\|^2 - \frac{2}{L}(f(w^t) - f(w^\star)) - \frac{\mu}{L}\|w^t - w^\star\|^2 + \frac{1}{L^2}\|\nabla f(w^t)\|^2 \\
&\leq (1 - \frac{\mu}{L})\|(w^t - w^\star)\|^2 - \frac{2}{L}(f(w^t) - f(w^\star)) + \frac{1}{L^2}\|\nabla f(w^t)\|^2 \\
&\overset{(1.2.8)}{\leq} (1 - \frac{\mu}{L})\|(w^t - w^\star)\|^2
\end{aligned}$$

□

2.2 Gradient Descent if f convex and L-smooth.

Lemma 2.1: Co-Coercivity

If f is convex and L-smooth:

$$1)\ f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2 \tag{2.2.1}$$

$$2)\ \langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_2 \tag{2.2.2}$$

Proof.

$$f(y) - f(x) = f(y) - f(z) + f(z) - f(x)$$

Convexity $\implies$

$$f(y) - f(z) \leq \langle \nabla f(y), y - z\rangle$$

Smoothness $\implies$

$$f(z) - f(x) \leq \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|^2$$

Therefore

$$
\begin{aligned}
f(y) - f(x) &\leq \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2}\|z - x\|^2 \\
&\leq \text{RHS}(z)
\end{aligned}
$$

We search for z which maximize RHS, $\nabla\text{RHS}(z) = 0$

$$
\begin{aligned}
\nabla\text{RHS}(z) = \nabla f(x) - \nabla f(y) + L(z - x) = 0 \\
\Leftrightarrow z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y))
\end{aligned}
$$

$$
\begin{aligned}
f(y) - f(x) &\leq \langle \nabla f(y), y - (x - \frac{1}{L}(\nabla f(x) - \nabla f(y))) \rangle \\
&\quad + \langle \nabla f(x), x - \frac{1}{L}(\nabla f(x) - \nabla f(y)) - x \rangle \\
&\quad + \frac{L}{2}\|x - \frac{1}{L}(\nabla f(x) - \nabla f(y)) - x\|^2 \\
&\leq \langle \nabla f(y), y - x \rangle + \frac{1}{L}\langle \nabla f(y), \nabla f(x) - \nabla f(y) \rangle \\
&\quad - \frac{1}{L}\langle \nabla f(x), \nabla f(x) - \nabla f(y) \rangle \\
&\quad + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \langle \nabla f(y), y - x \rangle - \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \\
&\quad + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2
\end{aligned}
$$

With $x \leftrightarrow y$, we prove (**??**). □

Theorem 2.2: Convergence GD 2

Let f be convex and L-smooth.

$$
f(w^t) - f(w^\star) \leq \frac{2L\|w^1 - w^\star\|_2^2}{t - 1} = O\left(\frac{1}{t}\right) \tag{2.2.3}
$$

where $w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$ for $t = 1, \ldots, T$.

$$
\implies \quad \text{for } \frac{f(w^T) - f(w^\star)}{\|w^1 - w^\star\|_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)
$$

Proof. With (**??**) and $\nabla f(w^\star) = 0$

$$
\begin{aligned}
\langle \nabla f(w^t), w^t - w^\star \rangle &\geq \frac{1}{L}\|\nabla f(w^t)\|_2 \\
-\langle \nabla f(w^t), w^t - w^\star \rangle &\leq \frac{1}{L}\|\nabla f(w^t)\|_2
\end{aligned} \tag{2.2.4}
$$

$$
\begin{aligned}
\|w^{t+1} - w^\star\|^2 &= \|w^t - w^\star - \frac{1}{L}\nabla f(w^t)\|^2 \\
&= \|w^t - w^\star\|^2 - \frac{2}{L}\langle \nabla f(w^t), w^t - w^\star \rangle + \frac{1}{L^2}\|\nabla f(w^t)\|^2 \\
&\overset{(??)}{\leq} \|w^t - w^\star\|^2 - \frac{1}{L^2}\|\nabla f(w^t)\|^2
\end{aligned}
$$

Therefore w^t converges.

$$\begin{aligned} f(w^t) - f(w^\star) &\overset{f \text{ is convex}}{\leq} \langle \nabla f(w^t), w^t - w^\star \rangle \\ &\leq \|\nabla f(w^t)\| \|w^t - w^\star\| \\ &\overset{w^t \text{ converges}}{\leq} \|\nabla f(w^t)\| \|w^1 - w^\star\| \\ \|\nabla f(w^t)\| &\geq \frac{f(w^t) - f(w^\star)}{\|w^1 - w^\star\|} \end{aligned} \tag{2.2.5}$$

f is L-smooth then

$$\begin{aligned} f(w^{t+1}) &\overset{(1.2.7)}{\leq} f(w^t) - \frac{1}{2L}\|\nabla f(w^t)\|^2 \\ f(w^{t+1}) - f(w^\star) &\leq f(w^t) - f(w^\star) - \frac{1}{2L}\|\nabla f(w^t)\|^2 \\ &\overset{(??)}{\leq} f(w^t) - f(w^\star) - \frac{1}{2L}\frac{(f(w^t) - f(w^\star))^2}{\|w^1 - w^\star\|^2} \end{aligned}$$

Let $\delta_t = f(w^t) - f(w^\star)$ and $C = \dfrac{1}{2L\|w^1 - w^\star\|^2}$ then

$$\begin{aligned} \delta_{t+1} &= \delta_t - C\delta_t^2 \\ \delta_{t+1}\frac{1}{\delta_{t+1}\delta_t} &= (\delta_t - C\delta_t^2)\frac{1}{\delta_{t+1}\delta_t} \\ \frac{1}{\delta_t} &= \frac{1}{\delta_{t+1}} - C\frac{\delta_t}{\delta_{t+1}} \\ C\frac{\delta_t}{\delta_{t+1}} &\leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \end{aligned}$$

$\dfrac{\delta_t}{\delta_{t+1}} \geq 1$

$$C \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}$$

Summing $t = 1, \ldots, T-1$

$$\frac{1}{\delta_T} - \frac{1}{\delta_1} \geq (T-1)C$$

$$\frac{1}{\delta_T} \geq (T-1)C \quad \text{because } \tfrac{1}{\delta_1} \geq 0$$

$$\delta_T \leq \frac{1}{(T-1)C}$$

□

2.3 Acceleration and lower bouds.

2.3.1 The Accelerated gradient method

Algorithm 1: Accelerated gradient

Set $w^1 = 0 = y^1, \kappa = L/\mu$
For $t = 1, 2, 3, \ldots, T$
. $\quad y^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$
. $\quad w^{t+1} = \left(1 + \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) y^{t+1} - frac\sqrt{\kappa} - 1\sqrt{\kappa} + 1w^t$

Output w^{T+1}

2.3.2 Convergence lower bounds strongly convex

Theorem 2.3: Nesterov 1

For any optimization algorithm where:

$$w^{t+1} \in w^t + span(\nabla f(w^1), \nabla f(w^2), \dots, \nabla f(w^t))$$

There exists a function $f(w)$ that is L-smooth and μ-strongly convex such that

$$\begin{aligned} f(w^T) - f(w^\star) &\geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa + 1}}\right)^{2(T-1)} \|w^1 - w^\star\|_2^2 \\ &= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right) \end{aligned} \tag{2.3.1}$$

2.3.3 Convergence lower bounds convex

Theorem 2.4: Nesterov 2

For any optimization algorithm where:

$$w^{t+1} \in w^t + span(\nabla f(w^1), \nabla f(w^2), \dots, \nabla f(w^t))$$

There exists a function $f(w)$ that is L-smooth and convex such that

$$\begin{aligned} \min_{i=1,\dots,T} f(w^i) - f(w^\star) &\geq \frac{3L\|w^1 - w^\star\|_2^2}{32(T+1)^2} \\ &= O\left(\frac{1}{T^2}\right) \end{aligned} \tag{2.3.2}$$

3 Proximal Operator and Methods

3.1 Proximal Operator

Definition 3.1: Training problem

$$w^\star = \arg\min_{w \in \mathbb{R}^d} L(w) + \lambda R(w)?$$

Definition 3.2: proximal operator

$$\mathrm{prox}_{\gamma R}(y) := \arg\min_{w} \frac{1}{2}\|w - y\|_2^2 + \gamma R(w) \tag{3.1.1}$$

Definition 3.3: subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n : f(y) \geq f(w) + \langle g, y - w\rangle, \forall y \in \mathrm{dom}(f)\} \tag{3.1.2}$$

We have

$$w^\star = \arg\min_{w} f(w) \Leftrightarrow 0 \in \partial f(w^\star) \tag{3.1.3}$$

Theorem 3.1: Assumptions

Assumptions for this class :
1. $L(w)$ is differentiable, $\mathcal{L}$-smooth and convex.
2. $R(w)$ is convex and "easy to optimize", i.e. $prox_{\gamma R}(y)$ is easy to find.

Lemma 3.1: Optimality Conditions

$$w^\star = \arg\min_{w \in \mathbb{R}^d} L(w) + \lambda R(w) \Leftrightarrow 0 \in \partial(L(w^\star) + \lambda R(w^\star)) \tag{3.1.4}$$

$$\Leftrightarrow -\nabla L(w^\star) \in \lambda \partial R(w^\star) \tag{3.1.5}$$

Theorem 3.2: proximal operator equivalence

Let f be a convex fonction. The proximal operator is

$$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v) \tag{3.1.6}$$

Proof.

$$\mathrm{prox}_f(v) := \arg\min_{w} \frac{1}{2}\|w - v\|_2^2 + f(w)$$

Let $w_v := \mathrm{prox}_f(v)$. Using optimality conditions:

$$0 \in \partial(\frac{1}{2}\|w_v - v\|^2 + f(w_v)) = w_v - v + \partial f(w_v)$$

Rearranging

$$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v)$$

□

Theorem 3.3: fixed point

Let $\min_{w} L(w) + \lambda R(w)$ be the training problem

$$w^\star = \mathrm{prox}_{\lambda\gamma R}(w^\star - \gamma \nabla L(w^\star)) \tag{3.1.7}$$

Optimal is a fixed point

Proof. Using (??)

$$\begin{aligned} w^\star = \arg\min_{w\in\mathbb{R}^d} L(w) + \lambda R(w) &\Leftrightarrow 0 \in \partial(L(w^\star) + \lambda R(w^\star)) \\ &\Leftrightarrow -\nabla L(w^\star) \in \lambda \partial R(w^\star) \\ &\Leftrightarrow w^\star + \gamma \nabla L(w^\star) \in w^\star - \gamma\lambda\partial R(w^\star) \\ &\Leftrightarrow w^\star \in (w^\star - \gamma\nabla L(w^\star)) - \gamma\lambda\partial R(w^\star) \\ &\overset{(??)}{\Leftrightarrow} w^\star = \mathrm{prox}_{\lambda\gamma R}(w^\star - \gamma\nabla L(w^\star)) \end{aligned}$$

$\square$

Proposition 3.1: Proximal Operator Properties

1. If $l(y, w) := f(y) + \langle \nabla f(y), w - y\rangle$ then $\mathrm{prox}_{\gamma L(y,.)}(y) = y - \gamma\nabla f(y)$
2. If $f(w) = \sum_{i=1}^{d} f_i(w_i)$ then $\mathrm{prox}_f(v) = (\mathrm{prox}_{f_1}(v_1), \ldots, \mathrm{prox}_{f_d}(v_d))$
3. If $f(w) = I_C(w) := \begin{cases} 0 & \text{if } w \in C \\ \infty & \text{if } w \notin C \end{cases}$ where C is closed and convex then $\mathrm{prox}_f(v) = \mathrm{proj}_C(v)$
4. If $f(w) = \langle b, w\rangle + c$ then $\mathrm{prox}_f(v) = v - b$
5. If $f(w) = \frac{\lambda}{2} w^T A w + \langle b, w\rangle$ where $A \succeq 0, A = A^T, \lambda \geq 0$ then $\mathrm{prox}_f(v) = (I + \lambda A)^{-1}(v - b)$
6. If $f(x) = \frac{1}{2}\|x\|_2^2$ then $\mathrm{prox}_{\lambda f}(x) = \frac{1}{1+\lambda}x$ (shrinkage operator)

Exercise 3.2.0. Let

$$l(y, w) := f(y) + \langle \nabla f(y), w - y\rangle$$

Show that

$$\mathrm{prox}_{\gamma L(y,.)}(y) = y - \gamma\nabla f(y)$$

i.e. A gradient step is also a proximal step.

Proof.

$$\begin{aligned} \mathrm{prox}_{\gamma L(y,.)}(y) &= \arg\min_{w} \frac{1}{2}\|w - y\|_2^2 + \gamma f(y) + \langle \gamma\nabla f(y), w - y\rangle \\ &= \arg\min_{w} \frac{1}{2}\|w - y\|_2^2 + \langle \gamma\nabla f(y), w - y\rangle \\ &= \arg\min_{w} \frac{1}{2}\|(w - y) + \gamma\nabla f(y)\|_2^2 \\ &= y - \gamma\nabla f(y) \end{aligned}$$

$\square$

Exercise 3.2.1. If

$$f(w) = \sum_{i=1}^{d} f_i(w_i) \text{ then } \mathrm{prox}_f(v) = (\mathrm{prox}_{f_1}(v_1), \ldots, \mathrm{prox}_{f_d}(v_d))$$

Proof. a faire

$$\text{prox}_f(v) = \arg\min_w \frac{1}{2}\|w-v\|_2^2 + \sum_{i=1}^d f_i(w_i)$$

$$\begin{aligned}\min_w \frac{1}{2}\|w-v\|_2^2 + \sum_{i=1}^d f_i(w_i) &= \min_w \frac{1}{2}\sum_{i=1}^d (w_i - v_i)^2 + \sum_{i=1}^d f_i(w_i) \\ &= \sum_{i=1}^d \min_w \frac{1}{2}(w_i - v_i)^2 + f_i(w_i)\end{aligned}$$

$$w_i^\star = \arg\min_w \frac{1}{2}(w_i - v_i)^2 + f_i(w_i) \Rightarrow w_i^\star = \text{prox}_{f_i}(v_i)$$

□

Exercise 3.2.2. If $f(w) = I_C(w) := \begin{cases} 0 & \text{if } w \in C \\ \infty & \text{if } w \notin C \end{cases}$ where C is closed and convex then $\text{prox}_f(v) = \text{proj}_C(v)$

Proof.

$$\text{prox}_f(v) = \arg\min_w \frac{1}{2}\|w-v\|_2^2 + f(w)$$

□

Exercise 3.2.3. If $f(w) = \langle b, w\rangle + c$ then $\text{prox}_f(v) = v - b$

Proof.

$$\text{prox}_f(v) = \arg\min_w \frac{1}{2}\|w-v\|_2^2 + f(w)$$

□

Exercise 3.2.4. If $f(w) = \frac{\lambda}{2}w^T Aw + \langle b, w\rangle$ where $A \succeq 0, A = A^T, \lambda \geq 0$ then $\text{prox}_f(v) = (I + \lambda A)^{-1}(v - b)$

Proof.

$$\text{prox}_f(v) = \arg\min_w \frac{1}{2}\|w-v\|_2^2 + f(w)$$

□

3.2 Thresholding

Proposition 3.2: Soft Thresholding

$$\text{prox}_{\lambda|.|}(x) = S_\lambda(x) := \text{sign}(x)(|x| - \lambda)_+ = \begin{cases} x - \lambda & \text{if } \lambda < x \\ 0 & \text{if } -\lambda \leq x \leq \lambda \\ x + \lambda & \text{if } \lambda > x \end{cases}$$

$$\text{prox}_{\lambda\|.\|_1}(v) = [S_\lambda(v_1), \ldots, S_\lambda(v_d)] := S_\lambda(v) = \text{sign}(v) \odot (|v| - \lambda)_+ \tag{3.2.1}$$

Proof. Let $\alpha \in \mathbb{R}$. If $\alpha^\star = \arg\min_\alpha \frac{1}{2}(\alpha - v)^2 + \lambda|\alpha|$ then (**??**) $\alpha^\star \in v - \lambda\partial|\alpha^\star|$

$$\alpha^\star \in \begin{cases} v - \lambda & \text{if } \alpha^\star > 0 \\ 0 & \text{if } \alpha^\star = 0 \\ v + \lambda & \text{if } \alpha^\star < 0 \end{cases}$$

$S_\lambda(v) = [S_\lambda(v_1), \ldots, S_\lambda(v_d)]$ using separability of $\|.\|_1$ and exercise 3.2.1. □

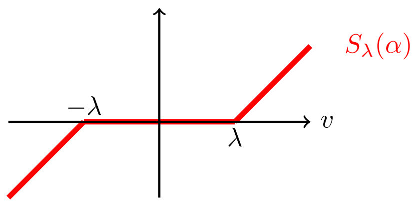

Figure 2: Soft Thresholding

Definition 3.4: Nuclear Norm

$$\|W\|_\star := \sum_{i=1}^{d} |\sigma_i(w)|$$

Definition 3.5: Frobenius Norm

$$\|A\|_F^2 := \mathrm{Tr}(A^T A)$$

Proposition 3.3: Invariance of $\|.\|_F$ and $\|.\|_\star$ under rotation

For any matrix A and orthogonal matrices O and Q

$$\|A\|_F^2 = \|OA\|_F^2 = \|AQ\|_F^2$$

$$\|A\|_\star = \|OA\|_\star = \|AQ\|_\star$$

Proof.

$$Tr((OA)^T OA) = Tr(A^T O^T OA) = Tr(A^T A)$$

If $A = U\mathrm{diag}(\sigma_i(A))V^T$:

$$OA = OU\mathrm{diag}(\sigma_i(A))V^T$$

which is the SVD of OA. Therefore OA and A have the same singular values. □

Theorem 3.4: Von Neumann 1937

For any matrix X and A of the same dimensions and orthogonal matrix U and V,

$$\langle UXV^T, A\rangle \leq \langle \mathrm{diag}(\sigma_i(X)), \mathrm{diag}(\sigma_i(A))\rangle$$

Definition 3.6: Extension of proximal operator to matrices

$$\mathrm{prox}_F(A) := \arg\min_{X\in\mathbb{R}^{d\times d}} \frac{1}{2}\|W - A\|_F^2 + F(X)$$

Proposition 3.4: Singular Value Thresholding

$$\mathrm{prox}_{\lambda\|.\|_\star}(A) := \arg\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{2}\|W - A\|_F^2 + \lambda\|W\|_\star = US_\lambda(\mathrm{diag}(\sigma(A)))V^T$$

where $A = U\mathrm{diag}(\sigma(A))V^T$ is a SVD decomposition.

Proof. Using proposition 3.3.

$$\begin{aligned}\frac{1}{2}\|W-A\|_F^2+\lambda\|W\|_\star &= \frac{1}{2}\|U^T(W-A)V\|_F^2+\lambda\|U^TWV\|_\star \\ &= \frac{1}{2}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\overline{W}\|_\star \qquad \text{with } \overline{W}=U^TWV\end{aligned}$$

Let $W=O\mathrm{diag}(\sigma(W))Q^T$ be the SVD of W. $\overline{W}=U^TO\mathrm{diag}(\sigma(W))Q^TV$

$$\begin{aligned}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2 &= \|\overline{W}\|_F^2+\|\mathrm{diag}(\sigma_i(A))\|_F^2-2\langle\overline{W},\mathrm{diag}(\sigma_i(A))\rangle \\ &= \|\mathrm{diag}(\sigma_i(W))\|_F^2+\|\mathrm{diag}(\sigma_i(A))\|_F^2-2\langle U^TO\mathrm{diag}(\sigma(W))Q^TV,\mathrm{diag}(\sigma_i(A))\rangle \\ &\overset{\text{th von Neuman}}{\geq} \|\mathrm{diag}(\sigma_i(W))\|_F^2+\|\mathrm{diag}(\sigma_i(A))\|_F^2-2\langle,\mathrm{diag}(\sigma_i(X))\mathrm{diag}(\sigma_i(A))\rangle \\ &\geq \|\mathrm{diag}(\sigma_i(W))-\mathrm{diag}(\sigma_i(A))\|_F^2\end{aligned}$$

$$\begin{aligned}\min_{W\in\mathbb{R}^{d\times d}}\frac{1}{2}\|W-A\|_F^2+\lambda\|W\|_\star &= \min_W\frac{1}{2}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\overline{W}\|_\star \\ &= \min_{\overline{W}}\frac{1}{2}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\overline{W}\|_\star \\ &= \min_{\overline{W}}\frac{1}{2}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\mathrm{diag}(\sigma_i(\overline{W}))\|_\star \\ &\geq \min_{\overline{W}}\frac{1}{2}\|\mathrm{diag}(\sigma_i(W))-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\mathrm{diag}(\sigma_i(\overline{W}))\|_\star \\ &= \min_{\overline{W}}\frac{1}{2}\|\mathrm{diag}(\sigma_i(\overline{W}))-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\mathrm{diag}(\sigma_i(\overline{W}))\|_\star\end{aligned} \tag{3.2.2}$$

Therefore the solution $\overline{W}$ will be a diagonal matrix.
Let $\overline{W}=\mathrm{diag}(\overline{W}_{ii})$, and $\overline{w}=(\overline{W}_{11},\dots,\overline{W}_{dd})$ be the vectorization of $\overline{W}$.
Thus $\|\overline{W}\|_\star=\|\overline{w}\|_1$ and $\|\overline{W}\|_F^2=\|\overline{w}\|_2^2$
Finally (**??**) becomes

$$\begin{aligned}\min_{W\in\mathbb{R}^{d\times d}}\frac{1}{2}\|W-A\|_F^2+\lambda\|W\|_\star &= \min_W\frac{1}{2}\|\overline{W}-\mathrm{diag}(\sigma_i(A))\|_F^2+\lambda\|\overline{W}\|_\star \\ &= \min_{\overline{w}\in\mathbb{R}^d}\frac{1}{2}\|\overline{w}-\mathrm{diag}(\sigma(A))\|_2^2+\lambda\|\overline{w}\|_1\end{aligned}$$

Consequently

$$ffff$$

□

3.3 Proximal Method

3.3.1 Proximal Method

Using $\mathcal{L}$- smoothness of L:

$$L(w)\leq L(y)+\langle\nabla L(y),w-y\rangle+\frac{\mathcal{L}}{2}\|w-y\|^2 \quad \forall w,y\in\mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent :

$$w=y-\frac{1}{\mathcal{L}}\nabla L(y)$$

But what about $R(w)$? Adding on $\lambda R(w)$ to upper bound:

$$L(w)+\lambda R(w)\leq L(y)+\langle\nabla L(y),w-y\rangle+\frac{\mathcal{L}}{2}\|w-y\|^2+\lambda R(w)$$

Can we minimized the RHS?

$$\begin{aligned}
\arg\min RHS(w) &= \arg\min_w \quad L(y) + \langle \nabla L(y), w-y\rangle + \frac{\mathcal{L}}{2}\|w-y\|^2 + \lambda R(w) \\
&= \arg\min_w \quad \langle \nabla L(y), w-y\rangle + \frac{\mathcal{L}}{2}\|w-y\|^2 + \lambda R(w) \\
&= \arg\min_w \quad \langle \frac{1}{\mathcal{L}}\nabla L(y), w-y\rangle + \frac{1}{2}\|w-y\|^2 + \frac{\lambda}{\mathcal{L}} R(w) && [\text{divising by } \frac{1}{\mathcal{L}}] \\
&= \arg\min_w \quad \frac{1}{2}\|\frac{1}{\mathcal{L}}\nabla L(y) - (w-y)\|^2 + \frac{\lambda}{\mathcal{L}} R(w) \\
&= \arg\min_w \quad \frac{1}{2}\|w - (y - \frac{1}{\mathcal{L}}\nabla L(y))\|^2 + \frac{\lambda}{\mathcal{L}} R(w) \\
&:= \operatorname{prox}_{\frac{\lambda}{\mathcal{L}}R}(y - \frac{1}{\mathcal{L}}\nabla L(y))
\end{aligned}$$

3.3.2 The Proximal Gradient Method

Solving the training problem $min_w L(w) + \lambda R(w)$
Where
1. $L(w)$ is differentiable, $\mathcal{L}$-smooth and convex.
2. $R(w)$ is convex and prox-friendly

Algorithm 2: Proximal Gradient Descent, ISTA

Set $w^1 = 0$
For $t = 1, 2, 3, \ldots, T$
. $\quad w^{t+1} = \textbf{prox}_{\lambda R/\mathcal{L}}\left(w^t - \frac{1}{\mathcal{L}}\nabla L(w^t)\right)$
Output w^{T+1}

Example: Lasso

$$\min_{w\in\mathbb{R}^d} \frac{1}{2n}\|Aw-y\|_2^2 + \lambda\|w\|_1$$

with $A = [a^1, \ldots, A^n]^T$, $\sum_{i=1}^n (y^i - \langle w, a^i\rangle)^2 = \|Aw-y\|_2^2$

$$\begin{aligned}
w^{t+1} &= \operatorname{prox}_{\lambda\|w\|_1/\mathcal{L}}\left(w^t - \frac{1}{n\mathcal{L}}A^T(Aw^t - y)\right) \\
&= S_{\lambda/\mathcal{L}}\left(w^t - \frac{1}{\sigma_{\max}(A)^2}A^T(Aw^t - y)\right) && [\mathcal{L} = \frac{\sigma_{\max}(A)^2}{n}, \text{ cf exercise 2.6}]
\end{aligned}$$

Theorem 3.5: Convergence of the Proximal Gradient Descent.

(Beck Teboulle 2009)
Let $f(w) = L(w) + \lambda R(w)$ where
1. $L(w)$ is differentiable, $\mathcal{L}$-smooth and convex.
2. $R(w)$ is convex and prox-friendly
Then

$$f(w^T) - f(w^\star) \le \frac{L\|w^1 - w^\star\|_2^2}{2T} = O\left(\frac{1}{T}\right) \tag{3.3.1}$$

where

$$w^{t+1} = \operatorname{prox}_{\lambda R/\mathcal{L}}\left(w^t - \frac{1}{\mathcal{L}}\nabla L(w^t)\right)$$

3.3.3 The FISTA Method

Algorithm 3: The FISTA Algorithm.

Set $w^1 = 0, z^1 = 0, \beta^1 = 1$
For $t = 1, 2, 3, \ldots, T$

. $w^{t+1} = \textbf{prox}_{\lambda R/\mathcal{L}}\left(z^t - \frac{1}{\mathcal{L}}\nabla L(z^t)\right)$

. $\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2}$

. $z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}}(w^{t+1} - w^t)$

Output w^{T+1}

Theorem 3.6: Convergence of FISTA.

(Beck Teboulle 2009)
Let $f(w) = L(w) + \lambda R(w)$ where
1. $L(w)$ is differentiable, $\mathcal{L}$-smooth and convex.
2. $R(w)$ is convex and prox-friendly
Then

$$f(w^T) - f(w^\star) \leq \frac{2L\|w^1 - w^\star\|_2^2}{(T+1)^2} = O\left(\frac{1}{T^2}\right) \tag{3.3.2}$$

where w^t are given by the FISTA algorithm

4 Stochastic Gradient Descent.

4.1 Solving the Finite Sum Training Problem

Definition 4.1: Datum Function

$$f_i(w) := l(h_w(x^i), y^i) + \lambda R(w)$$

Definition 4.2: Finite Sum Training Problem

$$f(w) := \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Can we use this sum structure?

Algorithm: Gradient Descent

Set $w^0 = 0$, **choose** $\alpha > 0$
For $t = 0, 1, 2, \dots, T$
. $\quad w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$
Output w^T

Problem with Gradient Descent: Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point...

Is it possible to design a method that uses only the gradient of a single data function $f_i(w)$ at each iteration?

Proposition 4.1: Unbiased Estimate

Let j be a random index sampled from $\{1, ..., n\}$ selected uniformly at random. Then:

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w) \tag{4.1.1}$$

Proof.

$$\nabla_j f_j(x) = \nabla f(x) + \epsilon_j \text{ , with } \mathbb{E}(\epsilon_j) = 0$$

□

Algorithm 4: SGD 0.0

Set $w^0 = 0$, **choose** $\alpha > 0$
For $t = 0, 1, 2, \dots, T-1$
. **sample** $j \in \{1, \dots, n\}$. $\quad w^{t+1} = w^t - \alpha \nabla f_j(w^t)$
Output w^T

Theorem 4.1: Convergence of SGD 0.0

If:

f is λ-strong convex
and (Expected Bounded Stochastic Gradients)

$$\mathbb{E}_j[\|\nabla f_j(w^t)\|_2^2] \le B^2, \forall w^t \text{ of SGD} \tag{4.1.2}$$

and $\frac{1}{\lambda} \ge \alpha > 0$
then the iterates of the SGD method satisfy:

$$\mathbb{E}[\|w^t - w^\star\|_2^2] \le (1-\alpha\lambda)^t \|w^0 - w^\star\|_2^2 + \frac{\alpha}{\lambda} B^2 \tag{4.1.3}$$

Proof.

$$\begin{aligned} f\lambda\text{-strong convex} &\implies f(y) \ge f(w) + \langle \nabla f(w), y - w\rangle + \frac{\lambda}{2}\|y-w\|_2^2, \forall w, y \\ &\implies f(w^\star) \ge f(w) + \langle \nabla f(w), w^\star - w\rangle + \frac{\lambda}{2}\|w^\star - w\|_2^2, \forall w \\ &\implies 2\langle \nabla f(w), w^\star - w\rangle \ge \lambda \|w^\star - w\|_2^2 + 2(f(w) - f(w^\star)) \\ &\implies 2\langle \nabla f(w), w^\star - w\rangle \ge \lambda \|w^\star - w\|_2^2 \end{aligned} \tag{4.1.4}$$

$$\begin{aligned} \|w^{t+1} - w^\star\|_2^2 &= \|w^t - w^\star - \alpha \nabla f_j(w^t)\|_2^2 \\ &= \|w^t - w^\star\|_2^2 - 2\alpha\langle \nabla f_j(w^t), w^t - w^\star\rangle + \alpha^2 \|\nabla f_j(w^t)\|_2^2 \end{aligned}$$

Taking expectation with respect to j, then using unbiaised estimator, bouded stochastic gradients and strong convexity:

$$\begin{aligned} \mathbb{E}_j[\|w^{t+1} - w^\star\|_2^2] &= \|w^t - w^\star\|_2^2 - 2\alpha\langle \mathbb{E}_j[\nabla f_j(w^t)], w^t - w^\star\rangle + \alpha^2 \mathbb{E}_j[\|\nabla f_j(w^t)\|_2^2] \\ &\overset{(2.1.1)}{=} \|w^t - w^\star\|_2^2 - 2\alpha\langle nabla f(w^t), w^t - w^\star\rangle + \alpha^2 \mathbb{E}_j[\|\nabla f_j(w^t)\|_2^2] \\ &\overset{(2.1.2)}{\le} \|w^t - w^\star\|_2^2 - 2\alpha\langle \nabla f(w^t), w^t - w^\star\rangle + \alpha^2 B^2 \\ &\overset{(2.1.4)}{\le} (1-\alpha\lambda)\|w^t - w^\star\|_2^2 + \alpha^2 B^2 \end{aligned}$$

Taking total expectation (law of total expectation) then summing up for 1 to T:

$$\begin{aligned} \mathbb{E}[\|w^{t+1} - w^\star\|_2^2] &\le (1-\alpha\lambda)\mathbb{E}[\|w^t - w^\star\|_2^2] + \alpha^2 B^2 \\ &\le (1-\alpha\lambda)^{t+1}\|w^0 - w^\star\|_2^2 + \sum_{i=0}^{t} (1-\alpha\lambda)^i \alpha^2 B^2 \end{aligned}$$

Using the geometric series sum $\displaystyle\sum_{i=0}^{t}(1-\alpha\lambda)^i = \frac{1-(1-\alpha\lambda)^{t+1}}{\alpha\lambda} \le \frac{1}{\alpha\lambda}$

$$\mathbb{E}[\|w^t - w^\star\|_2^2] \le (1-\alpha\lambda)^t \|w^0 - w^\star\|_2^2 + \frac{\alpha}{\lambda} B^2$$

□

4.2 SGD Shrinking stepsize

4.2.1 Shrinking SGD without average

Algorithm 5: SGD 1.1: Theorical

Set $w^1 = 0$, **choose** $\alpha_t \in \mathbb{R}_+$ **,** $\alpha_t \to 0$
For $t = 0, 1, 2, \ldots, T$
  sample $j \in \{1, \ldots, n\}$
  $w^{t+1} = \textbf{proj}_D(w^t - \alpha_t \nabla f_j(w^t))$
Output w^T

Theorem 4.2: Convergence of SGD 1.1 (Shrinking stepsize) - convex

If:
  f is convex
  and (Subgradients bounded)

$$\mathbb{E}_j[\|\nabla f_j(w^t)\|_2] \leq B, \forall w^t \text{ of SGD} \tag{4.2.1}$$

  and $\exists r \in \mathbb{R}_+ / w^\star \in D := w : \|w\| \leq r$
  and $\alpha_t = \dfrac{\alpha_0}{\sqrt{t+1}}$
then the iterates of the SGD 1.1 satisfy:

$$\mathbb{E}[f(w^T)] - f(w^\star) = O\left(\frac{1}{\sqrt{T}}\right) \tag{4.2.2}$$

(sublinear convergence)

Theorem 4.3: Convergence of SGD 1.2 (Shrinking stepsize) - strongly convex

If:
  f is λ-strongly convex
  and (Subgradients bounded) $\mathbb{E}_j[\|\nabla f_j(w^t)\|_2] \leq B, \forall w^t$ of SGD
  and $\exists r \in \mathbb{R}_+ / w^\star \in D := w : \|w\| \leq r$
  and $\alpha_t = \dfrac{\alpha_0}{\lambda t}$
then the iterates of the SGD 1.1 satisfy:

$$\mathbb{E}[f(w^T)] - f(w^\star) = O\left(\frac{1}{\lambda T}\right) \tag{4.2.3}$$

(faster sublinear convergence)

4.2.2 Shrinking SGD with average

Algorithm 6: SGDA 1.1 for Convex

Set $w^1 = 0$, **choose** $\alpha_t = \dfrac{2r}{B\sqrt{t}}$
For $t = 0, 1, 2, \ldots, T$
. **sample** $j \in \{1, \ldots, n\}$
. $w^{t+1} = \textbf{proj}_D(w^t - \alpha_t \nabla f_j(w^t))$
Output $\overline{w}^T = \dfrac{1}{T} \sum_{t=1}^{T} w^t$

Theorem 4.4: Convergence of SGDA 1.1 (Shrinking stepsize) - convex

If:

f is **convex**
and (Subgradients bounded) $\mathbb{E}_j[\|\nabla f_j(w^t)\|_2] \leq B, \forall w^t$ of SGD
and $\exists r \in \mathbb{R}_+ / w^\star \in D := w : \|w\| \leq r$

Let $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$

If $\alpha_t = \frac{\alpha_0}{\lambda t}$

then the iterates of the SGDA 1.1 satisfy:

$$\mathbb{E}[f(\overline{w}^T)] - f(w^\star) \leq \frac{3rB}{\sqrt{T}} \tag{4.2.4}$$

(sublinear convergence)

Proof.

$$\begin{aligned}
\|w^{t+1} - w^\star\|_2^2 &= \|w^t - w^\star - \alpha_t \nabla f_j(w^t)\|_2^2 \\
&= \|w^t - w^\star\|_2^2 - 2\alpha_t \langle \nabla f_j(w^t), w^t - w^\star \rangle + \alpha_t^2 \|\nabla f_j(w^t)\|_2^2
\end{aligned}$$

Taking expectation with respect to j, then using unbiaised estimator, bounded stochastic gradients and convexity:

$$\begin{aligned}
\mathbb{E}_j[\|w^{t+1} - w^\star\|_2^2] &= \|w^t - w^\star\|_2^2 - 2\alpha_t \langle \mathbb{E}_j[\nabla f_j(w^t)], w^t - w^\star \rangle + \alpha_t^2 \mathbb{E}_j[\|\nabla f_j(w^t)\|_2^2] \\
&\overset{(2.1.1)}{=} \|w^t - w^\star\|_2^2 - 2\alpha_t \langle nabla f(w^t), w^t - w^\star \rangle + \alpha_t^2 \mathbb{E}_j[\|\nabla f_j(w^t)\|_2^2] \\
&\overset{(2.1.2)}{\leq} \|w^t - w^\star\|_2^2 - 2\alpha_t \langle \nabla f(w^t), w^t - w^\star \rangle + \alpha_t^2 B^2 \\
&\overset{\text{convex.}}{\leq} \|w^t - w^\star\|_2^2 - 2\alpha_t (f(w^t) - f(w^\star)) + \alpha_t^2 B^2
\end{aligned}$$

Taking total expectation (law of total expectation) and re-arranging:

$$\mathbb{E}[\|w^{t+1} - w^\star\|_2^2] \leq \mathbb{E}_j[\|w^t - w^\star\|_2^2] - 2\alpha_t \mathbb{E}[f(w^t)] - 2\alpha_t f(w^\star)) + \alpha_t^2 B^2$$

$$\mathbb{E}[f(w^t)] - f(w^\star)) \leq \frac{1}{2\alpha_t}\mathbb{E}[\|w^t - w^\star\|_2^2] - \frac{1}{2\alpha_t}\mathbb{E}[\|w^{t+1} - w^\star\|_2^2] + \frac{\alpha_t}{2}B^2$$

Summing up for 1 to T:

$$\begin{aligned}
\sum_{t=1}^{T}(\mathbb{E}[f(w^t)] - f(w^\star)) \leq \frac{1}{2\alpha_1}\|w^1 - w^\star\|_2^2 + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\alpha t + 1} - \frac{1}{\alpha_t}\right)\mathbb{E}_j[\|w^t - w^\star\|_2^2] \\
- \frac{1}{2\alpha_{T+1}}\mathbb{E}[\|w^{T+1} - w^\star\|_2^2] + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t
\end{aligned}$$

Using $\|w\|^2 \leq r^2$ and $\alpha_{t+1} \leq \alpha_t$:

$$\begin{aligned}
\sum_{t=1}^{T}(\mathbb{E}[f(w^t)] - f(w^\star)) &\leq \frac{2r^2}{\alpha_1} + 2r^2\sum_{t=1}^{T-1}\left(\frac{1}{\alpha t + 1} - \frac{1}{\alpha_t}\right) + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t \\
&\leq \frac{2r^2}{\alpha_T} + \frac{B^2}{2}\sum_{t=1}^{T}\alpha_t
\end{aligned}$$

Let $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$ and divinding by T, using $\alpha_t = \frac{\alpha_0}{\sqrt{t}}$:

$$\begin{aligned}
\mathbb{E}[f(\overline{w}_T)] - f(w^\star) &\leq \frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[f(w^t)] - f(w^\star)) \\
&\leq \frac{r^2\sqrt{T}}{T\alpha_0} + \frac{B^2}{2T}\sum_{t=1}^{T}\frac{\alpha_0}{\sqrt{t}} \\
&\leq \frac{1}{\sqrt{T}}\left(\frac{2r^2}{\alpha_0} + \alpha_0 B^2\right)
\end{aligned}$$

Minimizing in α_0 gives $\alpha_0 = \sqrt{2}r/B$ thus:

$$\mathbb{E}[f(\overline{w}_T)] - f(w^\star) \leq \frac{3rB}{\sqrt{T}}$$

□

Algorithm 7: SGDA 1.2 for Strongly Convex

Set $w^0 = 0$, $\alpha_t = \dfrac{2}{\lambda(t+1)}$
For $t = 0, 1, 2, \ldots, T$
. **sample** $j \in \{1, \ldots, n\}$
. $w^{t+1} = \textbf{proj}_D(w^t - \alpha_t \nabla f_j(w^t))$
Output $\overline{w}^T = \dfrac{2}{T(T+1)}\sum_{t=0}^{T-1} tw^t$

Theorem 4.5: Convergence of SGDA 1.2 (Shrinking stepsize) - strongly convex

If:

f is λ-**strongly convex**
and (Subgradients bounded) $\mathbb{E}_j[\|\nabla f_j(w^t)\|_2] \leq B, \forall w^t$ of SGD
and $\exists r \in \mathbb{R}_+ / w^\star \in D := w : \|w\| \leq r$

Let $\overline{w}^T = \dfrac{2}{T(T+1)}\sum_{t=0}^{T-1} tw^t$

If $\alpha_t = \dfrac{2}{\lambda(t+1)}$

then the iterates of the SGDA 1.2 satisfies:

$$\mathbb{E}[f(\overline{w}^T)] - f(w^\star) \leq \frac{2B^2}{\lambda(T+1)} \tag{4.2.5}$$

(sublinear convergence)

4.3 Lazy SDG for Sparse Data

Consider the Finite Sum Training Problem with **L2 regularizor** and **linear hypothesis**:

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} l(\langle w, x^i\rangle, y^i) + \frac{\lambda}{2}\|w\|_2^2$$

Assume that **each data point** x^i **is** s**-sparse**, how many operations does each SGD step cost?

$$\begin{aligned} w^{t+1} &= w^t - \alpha_t(l'(\langle w^t, x^i\rangle, y^i)x^i + \lambda w^t) \\ &= \underbrace{(1-\lambda\alpha_t)w^t}_{\text{Rescaling } O(d)} - \underbrace{\alpha_t l'(\langle w^t, x^i\rangle, y^i)x^i}_{\text{+add sparse vector } O(s)=O(d)} \end{aligned}$$

Idea : rewrite the iterates using $w^t = \beta_t z^t$ where $\beta_t \in \mathbb{R}, z^t \in \mathbb{R}^d$:

$$\begin{aligned} \beta_{t+1}z^{t+1} &= (1-\lambda\alpha_t)\beta_t z^t - \alpha_t l'(\beta_t\langle z^t, x^i\rangle, y^i)x^i \\ &= \underbrace{(1-\lambda\alpha_t)\beta_t}_{\beta_{t+1}}\left(\underbrace{z^t - \frac{\alpha_t l'(\beta_t\langle z^t, x^i\rangle, y^i)}{(1-\lambda\alpha_t)\beta_t}x^i}_{z^{t+1}}\right) \end{aligned}$$

Each iteration is $O(s)$.

5 Variance Reduced Methods

References : [4], [2] [1]

5.1 Build an Estimate of the Gradient

Idea: Instead of using directly $\nabla f_j(w^t) \approx \nabla f(w^t)$, use $\nabla f_j(w^t)$ to estimate $g_t \approx \nabla f(w^t)$.
And the gradient step becomes $w^{t+1} = w^t - \alpha g^t$.

We would like **gradient estimate** such that:

Similar : $g^t \approx \nabla f(w^t)$ (typically unbiased $\mathbb{E}[g^t] = \nabla f(w^t)$)
Converges in L_2 : $\mathbb{E}\|g^t - \nabla f(w^t)\|_2^2 \underset{w^t \to w^*}{\to} 0$

Definition 5.1: Variance, Covariance

Let x and z be random variables.

$$\mathbb{VAR}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$$

$$\text{cov}(x, z) := \mathbb{E}[(x - \mathbb{E}[x])(z - \mathbb{E}[z])]$$

Definition 5.2: Covariates

Let x and z be random variables. We say that x and z are covariates if:

$$\text{cov}(x, z) \geq 0$$

Definition 5.3: Variance Reduced Estimate

$$x_z = x - z + \mathbb{E}[z]$$

Proposition 5.1: Variance Reduced Estimate Properties

1.
$$\mathbb{E}[x_z] = \mathbb{E}[x] \tag{5.1.1}$$

2.
$$\mathbb{VAR}[x_z] = \mathbb{VAR}[x] - 2\text{cov}(x, z) + \mathbb{VAR}[z] \tag{5.1.2}$$

5.2 Exercises

Exercise 1. Calculate L_i and $L_{\max} := \max_{i=1,\ldots n} L_i$ for

$$f(w) = \frac{1}{2}\|Aw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$$

Proof.

$$\begin{aligned}
f(w) &= \frac{1}{2}\|Aw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \\
&= \frac{1}{n}\left(\frac{n}{2}\sum_{i=1}^{n}(A_{i:}^T w - y_i)^2 + n\frac{\lambda}{2}\|w\|_2^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{2}(A_{i:}^T w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} f_i(w)
\end{aligned}$$

$$
\begin{aligned}
f_i(w) &= \frac{n}{2}(A_{i:}^T w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2 \\
\nabla f_i(w) &= nA_{i:}(A_{i:}^T w - y_i) + \lambda w
\end{aligned}
$$

$$
\begin{aligned}
\nabla^2 f_i(w) &= nA_{i:}A_{i:}^T + \lambda \\
&\preceq (n\|A_{i:}\|_2^2 + \lambda)I \\
&= L_i I
\end{aligned}
$$

□

Exercise 2. Calculate L_i and $L_{\max} := \max_{i=1,\dots n} L_i$ for

$$
f(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}\|w\|_2^2
$$

Proof.

$$
\begin{aligned}
f_i(w) &= \ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}\|w\|_2^2 \\
\nabla f_i(w) &= \frac{-y_i a_i e^{-y_i\langle w, a_i\rangle}}{1 + e^{-y_i\langle w, a_i\rangle}} + \lambda w
\end{aligned}
$$

$$
\begin{aligned}
\nabla^2 f_i(w) = &\frac{(y_i^2 a_i a_i^T e^{-y_i\langle w, a_i\rangle})\left(1 + e^{-y_i\langle w, a_i\rangle}\right) - (-y_i a_i e^{-y_i\langle w, a_i\rangle})(-y_i a_i^T e^{-y_i\langle w, a_i\rangle})}{\left(1 + e^{-y_i\langle w, a_i\rangle}\right)^2} + \lambda \\
&= y_i^2 a_i a_i^T \frac{e^{-y_i\langle w, a_i\rangle}}{\left(1 + e^{-y_i\langle w, a_i\rangle}\right)^2} + \lambda \\
&\preceq \left(\frac{y_i^2\|a_i\|_2^2}{4} + \lambda\right) I \\
&= L_i I
\end{aligned}
$$

□

Exercise 2. Let $f(w)$ be L-smooth and $f_i(w)$ be L_i-smooth for $i = 1, \dots, n$
Show that

$$
L \le \frac{1}{n}\sum_{i=1}^{n} L_i \le L_{\max} := \max_{i=1,\dots n} L_i
$$

Proof. From definition of $f_i(w)$ smoothness:

$$
\begin{aligned}
f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) &\le \frac{1}{n}\sum_{i=1}^{n} f_i(y) + \langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y), w - y\rangle + \frac{1}{2n}\sum_{i=1}^{n} L_i\|w - y\|_2^2 \\
&= f(y) + \langle \nabla f(y), w - y\rangle + \frac{1}{2n}\sum_{i=1}^{n} L_i\|w - y\|_2^2
\end{aligned}
$$

□

5.3 Stochastic Variance Reduced Gradients (SVGR)

Definition 5.4: SVGR

$$w^{t+1} = w^t - \alpha g^t$$
$$\text{Reference point } \widetilde{w} \in \mathbb{R}^d$$
$$\text{Sample } \nabla f_i(w^t), i \in \{1, \dots, n\}\text{uniformly}$$
$$\text{Grad Estimate } g^t = \nabla f_i(w^t) - \nabla f_i(\widetilde{w}) + \nabla f(\widetilde{w})$$
$$x_z = \quad x \quad - \quad z \quad + \quad \mathbb{E}[z]$$

Algorithm 8: Stochastic Variance Reduced Gradients (SVGR)

Set $w^0 = 0$**, chose** $\alpha > 0, m \in \mathbb{N}$
$\widetilde{w}^0 = w^0$
for $t = 0, 1, 2, \dots, T-1$
  calculate $\nabla f(\widetilde{w}^t)$
  $w^0 = \widetilde{w}^t$
  for $k = 0, 1, 2, \dots, m-1$
    sample $i \in \{1, \dots, n\}$
    $g^k = \nabla f_i(w^k) - \nabla f_i(\widetilde{w}^t) + \nabla f(\widetilde{w}^t)$
    $w^{k+1} = w^k - \alpha g^k$
  Option I: $\widetilde{w}^{t+1} = w^m$
  Option II: $\widetilde{w}^{t+1} = \frac{1}{m} \sum_{i=O}^{m-1} w^i$
Output $\widetilde{w}^T$

Theorem 5.1: Convergence of SVGR [3]

If:

$f(w)$ is λ-strongly convex,
$f_i(w)$ $L_{\max}$-smooth,

if

$$\alpha = \frac{1}{10L_{\max}} \text{ and } m = \frac{20L_{\max}}{\lambda}$$

Then:

$$\mathbb{E}[f(\tilde{w}^t)] - f(w^\star) \leq \left(\frac{7}{8}\right)^t (f(\tilde{w}^0) - f(w^\star))$$

NB1: need $O(L_{\max}/\lambda)$ inner iterations to have linear convergence.

NB2: in practice use $\alpha = 1/L_{\max}, m = n$

Proof.

$$\begin{aligned}
\|w^{k+1} - w^\star\|_2^2 &= \|w^k - w^\star - \alpha g^k\|_2^2 \\
&= \|w^k - w^\star\|_2^2 - 2\alpha\langle g^k, w^k - w^\star\rangle + \alpha^2\|g^k\|_2^2
\end{aligned}$$

Taking expectation with respect to j:

$$\begin{aligned}\mathbb{E}_j\left[\|w^{k+1}-w^\star\|_2^2\right] &= \mathbb{E}_j\left[\|w^k-w^\star\|_2^2 - 2\alpha\langle g^k, w^k-w^\star\rangle + \alpha^2\|g^k\|_2^2\right] \\ &= \|w^k-w^\star\|_2^2 - 2\alpha\langle \mathbb{E}_j\left[g^k\right], w^k-w^\star\rangle + \alpha^2\mathbb{E}_j\left[\|g^k\|_2^2\right] \\ &\overset{(3.1.1)\&(2.1.1)}{=} \|w^k-w^\star\|_2^2 - 2\alpha\langle \nabla f(w^k), w^k-w^\star\rangle + \alpha^2\mathbb{E}_j\left[\|g^k\|_2^2\right] \\ &\overset{convex.}{\leq} \|w^k-w^\star\|_2^2 - 2\alpha(f(w^k)-f(w^\star)) + \alpha^2\mathbb{E}_j\left[\|g^k\|_2^2\right]\end{aligned}$$

Must control :$\mathbb{E}_j\left[\|g^k\|_2^2\right]$

Lemma 5.1: Smoothness Consequence

$$\mathbb{E}_j\left[\|\nabla f_i(w)-\nabla f_i(w^\star)\|_2^2\right] \leq 2L_{\max}(f(w)-f(w^\star)) \tag{5.3.1}$$

Let $g_i(w) = f_i(w) - f_i(w^\star) - \langle\nabla f_i(w^\star), w-w^\star\rangle$, which is L_i-smooth.
Convexity of $f_i \implies g_i(w) \geq 0, \forall w$.
Using property of smoothness (1.2.7):

$$g_i(w) \overset{g_i\geq 0}{\geq} g_i(w) - g_i(w - \frac{1}{L_i}\nabla g_i(w)) \overset{(1.2.7)}{\geq} \frac{1}{2L_i}\|\nabla g_i(w)\|_2^2 \geq \frac{1}{2L_{\max}}\|\nabla g_i(w)\|_2^2$$

Inserting definition of $g_i(w)$:

$$\frac{1}{2L_{\max}}\|\nabla f_i(w)-\nabla f_i(w^\star)\|_2^2 \leq f_i(w) - f_i(w^\star) - \langle\nabla f_i(w^\star), w-w^\star\rangle$$

Taking expectation of i, we obtain (3.3.1).

Lemma 5.2: Smoothness Consequence 2

$$\mathbb{E}\left[\|g^k\|_2^2\right] \leq 4L_{\max}(f(w^k)-f(w^\star)) + 4L_{\max}(f(\tilde{w}^t)-f(w^\star)) \tag{5.3.2}$$

Hint: use
(1) $\mathbb{E}[\|X-\mathbb{E}[X]\|_2^2] \leq \mathbb{E}[\|X\|_2^2]$ with $X = \nabla f_i(w^\star) - \nabla f_i(\tilde{w}^t)$
(2) $\|a+b\|_2^2 \leq \|a\|_2^2 + \|b\|_2^2$
and (3.3.1) □

5.4 Stochastic Average Gradient unbiased version (SAGA)

Definition 5.5: SAGA

$$w^{t+1} = w^t - \alpha g^t$$
$$\text{Sample } \nabla f_i(w^t), i \in \{1,\ldots,n\}\text{uniformly}$$
$$\text{Grad Estimate } g^t = \nabla f_i(w^t) - \nabla f_i(w_i^t) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(w_j^t)$$
$$x_z = \quad x \quad - \quad z \quad + \quad \mathbb{E}[z]$$
$$\text{store gradient } \nabla f_i(w_i^{t+1}) = \nabla f_i(w^t), \nabla f_i(w_j^{t+1}) = \nabla f_i(w_j^t)\forall j \neq i$$

Disadvantage : store a $d\times n$ matrix...

Algorithm 9: Stochastic Average Gradient unbiased version (SAGA)

Set $w^0 = 0$, $w_i^0 = w^0$ **for** $i = 1,\ldots,n$, **chose** $\alpha > 0$
Setup table $[\nabla f_1(w_1^0),\ldots,\nabla f_n(w_n^0)] \in \mathbb{R}^{d\times n}$
for $t = 0,1,2,\ldots,T-1$

sample $i \in \{1, \dots, n\}$

$g^t = \nabla f_i(w^t) - \nabla f_i(w_i^t) + \dfrac{1}{n}\sum_{j=1}^{n} \nabla f_j(w_j^t)$

$w^{t+1} = w^t - \alpha g^t$

$\nabla f_i(w_i^{t+1}) = \nabla f_i(w^t)$

$\nabla f_j(w_j^{t+1}) = \nabla f_j(w_j^t) \qquad \forall j \neq i$

Output w^T

5.5 Stochastic Average Gradient - biased version (SAG)

Usual formulas

Theorem 5.2: Taylor expansion

$$f(a+h) = f(a) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(a) h^k + \frac{1}{p!} \int_0^1 (1-s)^p D^{n+1} f(a+hs) h^{p+1} \partial s \tag{5.5.1}$$

Gradients

$$\nabla x^T a = \nabla a^T x = a$$

$$\nabla a^T x b = a b^T$$

$$\nabla \|Ax+b\|^2 = 2A^T(Ax+b)$$

$$\nabla b^T x^T x c = x(bc^T + cb^T)$$

$$\nabla x^T A x = (A + A^T)x$$

$$\nabla b^T x^T A x c = D^T x b c^T + D x c b^T$$

$$\nabla (Ax+a)^T C (Bx+b) = A^T C (Bx+b) + B^T C^T (Ax+a)$$

Matrix

$$\|Aw - y\|_2^2 = \sum_{i=1}^{n} (A_{i:}^T w - y_i)^2$$

$$\|A\|_F^2 := Tr(A^T A) = \sum \|A_{i:}\|_2^2$$

$$A^T A = \sum A_{i:}^T A_{i:}$$

$$A^T b = \sum A_{i:}^T b_i$$

References

[1] R. M. Gower. Sketch and Project: Randomized Iterative Methods for Linear Systems and Inverting Matrices. *ArXiv e-prints*, Dec. 2016.

[2] R. M. Gower, P. Richtárik, and F. Bach. Stochastic Quasi-Gradient Methods: Variance Reduction via Jacobian Sketching. *ArXiv e-prints*, May 2018.

[3] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[4] M. Schmidt, N. Le Roux, and F. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. *ArXiv e-prints*, Sept. 2013.