# SD-TSIA 211
# Optimization for Machine Learning
# 4 February 2022

Paper documents are allowed (lecture notes, exercises and books)
Electronic devices are forbidden
You may write in English or in French
2h-exam

**Exercise 1** (Comparison of algorithms)**.**
We consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

where for all $i$, $f_i$ is a convex differentiable function with a $L$-Lipschitz gradient. We assume that $f_i$ has bounded subgradients. We denote $F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$ and assume that $F$ is $\mu$-strongly convex. We assume that for all $i$, computing $\nabla f_i$ costs $C_{\nabla f}$ operations. We denote $\bar{x}_k^\gamma = \dfrac{\sum_{l=0}^{k}\gamma_l x_l}{\sum_{j=0}^{k}\gamma_j}$. Below, $C$ denotes a constant independent of $k$ and $N$.
For each algorithm, choose in the list which step-size sequence should be taken, which convergence speed holds and how much each iteration costs.

Algorithms

A. $i_{k+1} \sim U(\{1,\dots,N\})$
$x_{k+1} = x_k - \gamma_k \nabla f_{i_{k+1}}(x_k)$

B. $G_k \in \partial F(x_k)$
$x_{k+1} = x_k - \gamma_k G_k$

C. $x_{k+1} = x_k - \gamma_k \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_k)$

Step size sequence

i. $\gamma_k = \frac{1}{L}$

ii. $\gamma_k = \gamma$, $\gamma$ small enough

iii. $\gamma_k = \frac{1}{\sqrt{k+1}}$

iv. $\gamma_k = k$

Convergence speed

1. $F(x_k) - F(x^*) \leq \frac{C}{k}$
2. $F(x_k) - F(x^*) \leq (1-\frac{\mu}{L})^k C$
3. $\mathbb{E}[F(\bar{x}_k^\gamma) - F(x^*)] \leq \frac{C\ln(k)}{\sqrt{k}}$
4. $F(\bar{x}_k^\gamma) - F(x^*) \leq \frac{C\ln(k)}{\sqrt{k}}$
5. $\mathbb{E}[F(x_k) - F(x^*)] \leq (1-\gamma\mu)^k C + C\gamma$

Cost of each iteration

a. $C_{\nabla f}$

b. $NC_{\nabla f}$

c. $(C_{\nabla f})^2$

**Exercise 2** (Support vector machines in primal space)**.**
Given $N$ data points $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, we consider the following support vector machine problem

$$\min_{v\in\mathbb{R}^d, \xi\in\mathbb{R}^N} \frac{\|v\|_2^2}{2} + c\sum_{i=1}^N \xi_i^2 \text{ such that } \forall i \in \{1,\dots,N\},\ 1 - y_i\langle v, x_i\rangle \le \xi_i \text{ and } \xi_i \ge 0\,. \quad (1)$$

Note the square in $\xi_i^2$, which makes the model different from the one presented in the tutorial session. For a given $v$, we consider the partial minimization in $\xi$ and denote

$$\begin{aligned} F(v) = \min_{\xi\in\mathbb{R}^N} \frac{\|v\|_2^2}{2} + c\sum_{i=1}^N \xi_i^2 \\ \forall i \in \{1,\dots,N\},\ 1 - y_i\langle v, x_i\rangle \le \xi_i \\ \forall i \in \{1,\dots,N\}, \xi_i \ge 0 \end{aligned}$$

1. Show that for all $v \in \mathbb{R}^d$,

$$F(v) = \frac{\|v\|_2^2}{2} + c\sum_{i=1}^N \max(0, 1 - y_i\langle v, x_i\rangle)^2$$

2. Show that $F$ has a unique minimizer.

   his is a continuous strongly convex function, hence existence is given by lower semi-continuity and coercivity and uniqueness by strict convexity.

3. Denote $h(z) = \max(0, z)^2$. Give the formula of $\nabla h(z)$ for $z \in \mathbb{R}$.

4. Denote $g_i(v) = \max(0, 1 - y_i\langle v, x_i\rangle)^2$. Compute $\nabla g_i(v)$.

5. Write the stochastic gradient algorithm for

$$\min_{v\in\mathbb{R}^d} F(v)$$

   Specify in particular the step-size you are choosing.

6. How many iterations would we need to get a point $v_k$ such that $\mathbb{E}[F(v_k)] - \inf F \le \epsilon$? You may use the notation of Exercise 1 and neglect logarithmic terms.

**Exercise 3** (Support vector machines in dual space)**.**
We consider again the optimization problem (1)

$$\min_{v\in\mathbb{R}^d, \xi\in\mathbb{R}^n} \frac{\|v\|_2^2}{2} + c\sum_{i=1}^N \xi_i^2 \text{ such that } \forall i \in \{1,\dots,N\},\ 1 - y_i\langle v, x_i\rangle \le \xi_i \text{ and } \xi_i \ge 0\,.$$

1. Write the Lagrangian function $L(v, \xi, \phi, \alpha)$ associated to Problem (1).

2. By writing the Karush-Kuhn-Tucker condition, show that the primal variables $v$ and $\xi$ can be reconstructed as functions of the dual variables $\phi, \alpha$.

This last result ensures that if we are able to solve the dual problem, then we also have a solution to the primal problem.

3. Show that a dual problem to (1) is given by

$$\max_{\phi \geq 0} -\frac{1}{2}\phi^T DKD\phi + \mathbf{1}^T\phi - \frac{1}{4c}\|\phi\|_2^2 \tag{2}$$

   where $K = (\langle x_i, x_j \rangle)_{i,j=1\dots n}$, $D = \mathrm{diag}(y_1 \dots y_n)$ and $\mathbf{1}^T = (1, \dots, 1)$.

Equivalently, we can write this optimization problem as

$$\min_{\phi \in \mathbb{R}^N} f(\phi) + g(\phi)$$

where $f(\phi) = \frac{1}{2}\phi^T DKD\phi - \mathbf{1}^T\phi + \frac{1}{4c}\|\phi\|_2^2$ and $g(\phi) = \iota_{[0,+\infty[^N}(\phi) = \begin{cases} 0 & \text{if } 0 \leq \phi_i, \forall i \\ +\infty & \text{if } \exists i \ : \ \phi_i < 0 \end{cases}$

4. Calculate the gradient $\nabla f(\phi)$ of $f$ as function of $\phi, D, K$.

5. Calculate the proximal operator of $g$.

6. Write the proximal gradient algorithm for the resolution of (2). What step size are you choosing?

7. How many iterations would we need to get $\phi_k$ such that $f(\phi_k) + g(\phi_k) - \inf f + g \leq \epsilon$?

**Exercise 4.**
What are the advantages and drawbacks of the algorithms presented in Exercise 2 and Exercise 3.

## Ex 1.

A. $i_{k+1} \sim U(\{1, \cdots, N\})$

$x_{k+1} = x_k - \gamma_k \nabla f_{i_{k+1}}(x_k)$ $\rightarrow$ SGD: $\gamma_k = \frac{1}{\sqrt{k+1}}$

$\mathbb{E}[F(\bar{x}_k^T) - F(x^*)] \le \frac{C \ln(k)}{\sqrt{k}}$

cost: $C_{of}$

B: $G_k \in \partial F(x_k)$ $\rightarrow$ GD: $\gamma_k = \frac{1}{\sqrt{k+1}}$, $F(x_k) - F(x^*) \le (1 - \frac{\mu}{L})^k C$

$x_{k+1} = x_k - \gamma_k G_k$.

cost: $N C_{of}$

△ C. $x_{k+1} = x_k - \gamma_k \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_k)$ $\rightarrow$ 单的梯度下降.

if: $\gamma_k = \frac{1}{L}$, $F(x_k) - F(x^*) \le \frac{C}{k}$

if: $\gamma_k = \frac{1}{\sqrt{k+1}}$, $F(x_k) - F(x^*) \le \frac{C \ln k}{\sqrt{k}}$

cost: $N C_{of}$.

## Ex 2.

1. $F(v) = \min_{\xi \in \mathbb{R}^N} \left( \frac{\|v\|^2}{2} + C \sum_{i=1}^{N} \xi_i^2 \right) = \frac{\|v\|^2}{2} + \min_{\xi \in \mathbb{R}^N} \left( C \sum_{i=1}^{N} \xi_i^2 \right)$

for term $\sum_{i=1}^{N} \xi_i^2$, $\sum_{i=1}^{N} \xi_i^2 \ge \sum_{i=1}^{N} \max^2(0, 1 - y_i \langle v, x_i \rangle)$

therefore, $\min_{\xi \in \mathbb{R}^N} C \sum_{i=1}^{N} \xi_i^2 = C \sum_{i=1}^{N} \max^2(0, 1 - y_i \langle v, x_i \rangle)$

and then $F(v) = \frac{\|v\|^2}{2} + C \sum_{i=1}^{N} \max^2(0, 1 - y_i \langle v, x_i \rangle)$.

2. To prove $F(v)$ exist a unique minimizer, we need to show $F(v)$ is coercive and strictly convex.

$\lim_{\|v\| \to +\infty} F(v) = \infty$ which indicates that $F(v)$ is coercive.

For the strictly convex: $F(v) = \frac{\|v\|^2}{2} + C \sum_{i=1}^{N} \max^2(0, 1 - y_i \langle v, x_i \rangle)$

where $\frac{\|v\|^2}{2}$ is a strictly convex function, and both $0^2$ and $(1-y_i\langle v, x_i\rangle)^2$ are strictly convex functions.

Therefore, $F(v)$ is a strictly convex function.

3. $h(z) = \max^2(0, z) = \begin{cases} z^2, & z>0 \\ 0, & z\le 0 \end{cases}$

$\nabla h(z) = \begin{cases} 2z, & z>0. \\ 0, & z\le 0. \end{cases}$

4. $g_i(v) = \max^2(0, 1-y_i\langle v, x_i\rangle) = h(1-y_i\langle v, x_i\rangle)$.

$\nabla g_i(v) = \nabla h(1-y_i\langle v, x_i\rangle)\cdot(-y_i x_i)$

$= \begin{cases} -2y_i(1-y_i\langle v, x_i\rangle)x_i, & 1-y_i\langle v, x_i\rangle > 0 \\ 0, & \text{otherwise} \end{cases}$

5. Generate $I_{k+1} \sim U(\{1, \cdots, N\})$

$v^{k+1} = v^k - \gamma_k \nabla g_{I_{k+1}}(v^k) \qquad \gamma_k = \frac{\gamma_0}{\sqrt{k}}$

6. $\mathbb{E}[F(v_k)] - \inf F \le \frac{1}{\sqrt{k}} \le \varepsilon. \qquad k \ge \frac{1}{\varepsilon^2}$

Ex 3.

1. $L(v, \xi, \phi, \alpha) = \left(\frac{\|v\|^2}{2} + C\sum_{i=1}^N \xi_i^2\right) + \sum_{i=1}^N \phi_i(1-y_i\langle v, x_i\rangle - \xi_i) - \sum_{i=1}^N \alpha_i\xi_i$

$\phi_i \ge 0. \qquad \alpha_i \ge 0$

2. First order primal feasibility:

$\nabla_v L = 0: \quad v = \sum_{i=1}^{N} \phi_i y_i x_i$

$\nabla_\xi L = 0: \quad \xi_i = \frac{1}{2C}(\phi_i + \alpha_i)$

Primal feasibility:

$1 - y_i \langle v, x_i \rangle - \xi_i \leq 0$

$\xi_i \geq 0$

Dual feasibility:

$\phi_i \geq 0$

$\alpha_i \geq 0$

Complementary slackness:

$\phi_i (1 - y_i \langle v, x_i \rangle - \xi_i) = 0$

$\alpha_i \xi_i = 0$

if $\alpha_i = 0$. $\xi_i \geq 0$. $\xi_i = \frac{1}{2C}\phi_i \Rightarrow \begin{cases} \alpha_i = 0 \\ \phi_i = 2C\xi_i \end{cases}$

if $\xi_i = 0$ $\alpha_i \geq 0$. $\alpha_i = -\phi_i \leq 0$. $\Rightarrow \begin{cases} \alpha_i = 0 \\ \phi_i = 0 \end{cases}$

3. Dual: $\max_{\phi \geq 0} \inf_{v,\xi} L(v, \xi, \phi, \alpha)$

From $\nabla_v L = 0$, we get that $v^* = \sum_{i=1}^{N} \phi_i y_i x_i$

$\nabla_\xi L = 0$ $\quad \xi_i = \frac{1}{2C}(\phi_i + \alpha_i)$

$= \frac{1}{2C}\phi_i$

$$L = \frac{\|v\|^2}{2} + C\sum_{i=1}^{N}\xi_i^2 + \sum_{i=1}^{N}\phi_i\left(1 - y_i\langle v, x_i\rangle - \xi_i\right)$$

$$= \frac{1}{2}\langle v, v\rangle + C\sum_{i=1}^{N}\left(\frac{1}{2C}\phi_i\right)^2 + \sum_{i=1}^{N}\phi_i - \sum_{i=1}^{N}\phi_i y_i\langle v, x_i\rangle - \sum_{i=1}^{N}\phi_i\xi_i$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_i\phi_j y_i y_j\langle x_i, x_j\rangle$$

$$+ \frac{1}{4C}\|\phi\|^2$$

$$+ 1^T\phi$$

$$- \sum_{i=1}^{N}\sum_{j=1}^{N}\phi_i y_i \phi_j y_j\langle x_i, x_j\rangle$$

$$- \frac{1}{2C}\|\phi\|^2$$

$$= -\frac{1}{2}\phi^T DKD\phi + 1^T\phi - \frac{1}{4C}\|\phi\|^2$$

Thus, (1) $\Leftrightarrow$ $\max_{\phi \geq 0} -\frac{1}{2}\phi^T DKD\phi + 1^T\phi - \frac{1}{4C}\|\phi\|^2$

4. $f(\phi+\Delta) = \frac{1}{2}(\phi+\Delta)^T DKD(\phi+\Delta) - 1^T(\phi+\Delta) + \frac{1}{4C}\|\phi+\Delta\|^2$

$$= \boxed{\frac{1}{2}\phi^T DKD\phi - 1^T\phi + \frac{1}{4C}\|\phi\|^2} = f(\phi)$$

$$+ \frac{1}{2}\Delta^T DKD\phi - 1^T\Delta + \frac{1}{4C}\|\Delta\|^2$$

$$+ \frac{1}{2}\phi^T DKD\Delta + \frac{1}{2C}\langle\phi, \Delta\rangle$$

$$+ \frac{1}{2}\Delta^T DKD\Delta \times$$

$$= f(\phi) + \langle DKD\phi/2, \Delta\rangle + \langle DKD\phi/2, \Delta\rangle$$

$$- \langle 1, \Delta\rangle + \frac{1}{2C}\langle\phi, \Delta\rangle + o(\Delta)$$

$$= f(\phi) + \langle DKD\phi - 1 + \frac{1}{2C}\phi, \Delta\rangle + o(\Delta)$$

Thus, $\nabla f(\phi) = DKD\phi - 1 + \frac{1}{2C}\phi$.

5. $\text{prox}_{\gamma g}(x) = \underset{y}{\text{argmin}}\ g(y) + \frac{1}{2}\|x-y\|^2$

$= \underset{y}{\text{argmin}}\ \gamma_{[0,+\infty)^N}(y) + \frac{1}{2}\|x-y\|^2.$

$= \underset{y \in [0,+\infty)^N}{\text{argmin}}\ \frac{1}{2}\|x-y\|^2.$

$= (x)_+$

6. $\phi^{k+1} = \text{prox}_{\gamma g}(\phi^k - \gamma_k \nabla f(\phi^k))$

$= \left(\phi^k - \gamma_k\left(DKD\phi^k - 1 + \frac{1}{2C}\phi^k\right)\right)_+$

$$\|\nabla f(\phi_1) - \nabla f(\phi_2)\| = \left\| \begin{matrix} \left(DKD\phi_1 - 1 + \frac{1}{2C}\phi_1\right) \\ -\left(DKD\phi_2 - 1 + \frac{1}{2C}\phi_2\right) \end{matrix} \right\|$$

$$= \|\left(DKD + \frac{1}{2C}Id\right)(\phi_1 - \phi_2)\| \leq \|DKD + \frac{1}{2C}Id\|\,\|\phi_1 - \phi_2\|$$

and shows that $\nabla f(\phi)$ is $L$-Lipschitz continuous with $L = \|DKD + \frac{1}{2C}Id\|$.

Therefore, we choose $\gamma_k = \frac{1}{L}$.

7. $f(\phi_k) + g(\phi_k) - \inf(f+g) \leq \frac{L\|\phi_0 - \phi^*\|^2}{2k} \leq \varepsilon$

$\Rightarrow k \geq \frac{L\|\phi_0 - \phi^*\|^2}{2\varepsilon}$

Ex 4.

advantage in 2: convergence faster than 3.
SGD only use one data in each iteration.

drawback in 2:

advantage in 3: fixed steps could reduce the cost of computation.

drawback in 3: