OPTIMIZATION FOR MACHINE LEARNING

November 21, 2024

Today:

→ Exercise solutions

→ Sparse optimization / ℓ_1 regularization

Exercise on stochastic gradient (from lecture 11)

Setup: $\underset{x \in \mathbb{R}^d}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

f_i depends on the ith sample of a dataset of size n

Batch SG

$$x_{k+1} = x_k - \frac{\alpha_k}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k)$$

S_k is a set of indices drawn randomly in $\{1, \dots, n\}$ with/without replacement

This exercise

$|S_k| = n_b \in \{1, \dots, n\} \quad \forall k.$

$$\forall S \subseteq \{1, \dots, n\}, \quad \mathbb{P}(S_k = S) = \begin{cases} 0 & \text{if } |S| \neq n_b \\ \frac{1}{\binom{n}{n_b}} & \text{if } |S| = n_b \end{cases}$$

→ Sampling n_b indices without replacement

→ Uniform sampling over all possible subsets of $\{1, \dots, n\}$ with cardinality n_b

a) Show $\mathbb{E}_{S_k}\left[\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k)\right] = \nabla f(x_k) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k)$

$$\mathbb{E}_{S_k}\left[\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k)\right] = \sum_{S \subseteq \{1, \dots, n\}} \mathbb{P}(S_k = S) \times \frac{1}{|S|} \sum_{i \in S} \nabla f_i(x_k)$$

$$= \sum_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = n_b}} \mathbb{P}(S_k = S) \times \frac{1}{n_b} \sum_{i \in S} \nabla f_i(x_k)$$

$$= \sum_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = n_b}} \frac{1}{\binom{n}{n_b}} \times \frac{1}{n_b} \sum_{i \in S} \nabla f_i(x_k)$$

$$= \frac{1}{\binom{n}{n_b} \times n} \sum_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = n_b}} \sum_{i \in S} \nabla f_i(x_k)$$

Given some index $j \in \{1, \dots, n\}$, how many times does $\nabla f_j(x_k)$ appear in the double sum?

$\binom{n-1}{n_b - 1}$

$n - 1 \to$ number of possibilities (sampling without replacement)

$n_b - 1 \to n_b - 1$ indices left to pick

$$\sum_{\substack{S \subseteq \{1, \dots, n\} \\ |S| = n_b}} \sum_{i \in S_k} \nabla f_i(x_k) = \sum_{i=1}^{n} \binom{n-1}{n_b - 1} \nabla f_i(x_k)$$

Hence $$\mathbb{E}_{S_k}\left[\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k) \right] = \frac{1}{\binom{n}{n_b} \times n_b} \sum_{i=1}^{n} \binom{n-1}{n_b - 1} \nabla f_i(x_k)$$

$$= \frac{\binom{n-1}{n_b - 1}}{\binom{n}{n_b} \times n_b} \sum_{i=1}^{n} \nabla f_i(x_k)$$

$$\binom{n}{n_b} = \frac{n}{n_b} \binom{n-1}{n_b - 1}$$

$$\binom{n}{n_b} = \frac{n!}{n_b!\,(n - n_b)!}$$

$$= \binom{n-1}{n_b - 1} \times \frac{1}{\frac{n}{n_b} \binom{n-1}{n_b - 1} \times n_b} \sum_{i=1}^{n} \nabla f_i(x_k)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k)$$

b) How can we guarantee that $\mathbb{E}[f(x_k) - f^*] \xrightarrow[k \to \infty]{} 0$ under the assumptions that f is $C_L^{1,1}$, μ-strongly convex and $f^* = \min_x f(x)$?

①. 1 variant of the proposed method (batch SG)

②. 1 modification to the algorithm

① Use a decaying stepsize

Other option: Iterate averaging (not covered in class) but see subgradient lecture

$n_b = n$

② Any gradient aggregation method (SAGA, SVRG)

Exercise from the previous lecture

$$(P) \quad \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \|x\|_1 + \frac{1}{2\alpha} \|x - u\|_2^2$$

for some $u \in \mathbb{R}^d$ and some $\alpha > 0$.

1) Write the solution of (P) as a proximal operator calculation

$$\mathrm{prox}_h(u) = \underset{x}{\mathrm{argmin}} \left\{ h(x) + \frac{1}{2}\|x-u\|_2^2 \right\}$$

$$\underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \|x\|_1 + \frac{1}{2\alpha}\|x-u\|_2^2 \right\}$$

$$= \underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \alpha\|x\|_1 + \frac{1}{2}\|x-u\|_2^2 \right\} = \mathrm{prox}_{\alpha\|\cdot\|_1}(u)$$

2) Write down the proximal gradient iteration for problem (P) with $f(x) = \frac{1}{2}\|x-u\|_2^2$ (data-fitting term)

$\Omega(x) = \|x\|_1$ (regularization)

$\lambda = \alpha$

Using the equivalence between (P) and

$$\underset{x \in \mathbb{R}^d}{\mathrm{minimize}} \ \frac{1}{2}\|x-u\|_2^2 + \alpha\|x\|_1 = f(x) + \lambda\Omega(x)$$

we write the proximal gradient iteration

$$x_{k+1} = \mathrm{prox}_{\lambda\alpha_k\Omega(\cdot)}\left(x_k - \alpha_k \nabla f(x_k)\right)$$

$$= \underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T(x-x_k) + \frac{1}{2\alpha_k}\|x-x_k\|_2^2 + \lambda\Omega(x) \right\}$$

$$= \underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \underbrace{\frac{1}{2}\|x_k-u\|^2}_{\text{constant w.r.t. } x} + (x_k-u)^T(x-x_k) + \frac{1}{2\alpha_k}\|x-x_k\|_2^2 + \alpha\|x\|_1 \right\}$$

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ (x_k - u)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 + \alpha \|x\|_1 \right\}$$

NB: If $x_k = u$ and $\alpha_k = 1$, the subproblem is the original problem

1 Proximal gradient and ℓ_1 regularization

Problem:

$$(P_1) \quad \underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \lambda \|x\|_1$$

where $f : \mathbb{R}^d \to \mathbb{R}$ $C_L^{1,1}$ $\left(f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \right)$

$$\|x\|_1 = \sum_{j=1}^d |x_j|$$

The function $f + \lambda \|\cdot\|_1$ is nonsmooth (because $\|\cdot\|_1$ is) and has a composite structure $\underbrace{\text{smooth}}_{f} + \underbrace{\text{nonsmooth}}_{\lambda\|\cdot\|_1}$.

In that case, the subdifferential of $f + \lambda\|\cdot\|_1$ at x is given by $\partial(f + \|\cdot\|_1)(x) = \{ v = \nabla f(x) + g \mid g \in \partial(\lambda \|\cdot\|_1)(x) \}$

Proposition: If x^* is a solution of (P_1), then

$$0 \in \partial(f + \lambda\|\cdot\|_1)(x^*) \,..$$

Counter-ex $d=1$

$f(x) = -x^2 \quad d=1$

$0 \in \partial(f + |\cdot|)(0)$ $\quad -x^2+|x|$

$$\iff \quad -\nabla f(x^0) \in \partial(\lambda \|\cdot\|_1)(x^0)$$

Remark : • The condition $-\nabla f(x^0) \in \partial(\lambda\|\cdot\|_1)(x^0)$ ("variational inequality") cannot be solved in general, even for simple fs

Ex) $f(x) = \frac{1}{2n}\|Ax-y\|^2$

• But within a proximal subproblem, we can actually solve the optimality condition.

↳ Suppose we apply the proximal gradient method to (P_1).

At iteration k, we need to solve

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \Big\{ \underbrace{f(x_k) + \nabla f(x_k)^T(x-x_k) + \frac{1}{2\alpha_k}\|x-x_k\|^2 + \lambda\|x\|_1}_{\varphi_k(x)} \Big\} \quad \alpha_k > 0$$

φ_k convex, even strongly convex and quadratic (and the quadratic term is simple), $C^{1,1}_{1/\alpha_k}$

$\forall x \in \mathbb{R}^d$, $\nabla \varphi_k(x) = \nabla f(x_k) + \frac{1}{\alpha_k}(x-x_k)$

For that problem, since $\varphi_k + \lambda\|.\|_1$ is convex,

$$x^* \text{ is solution of } (P_1) \iff 0 \in \partial\left(\varphi_k + \lambda\|.\|_1\right)(x^*)$$

$$\iff -\nabla\varphi_k(x^*) \in \partial\left(\lambda\|.\|_1\right)(x^*)$$

$$\iff \left[-\nabla f(x_k) - \frac{1}{\alpha_k}(x^* - x_k)\right] \in \partial\left(\lambda\|.\|_1\right)(x^*)$$

$\underline{d=1}$ $\quad t \mapsto |t| \qquad \partial(|.|)(t) = \begin{cases} 1 & \text{if } t>0 \\ -1 & \text{if } t<0 \\ [-1,1] & \text{if } t=0 \end{cases}$

$$\|x\|_1 = \sum_{j=1}^{d} |x_j|$$

$$\forall x \in \mathbb{R}^d, \ \partial\left(\|.\|_1\right)(x) = \left\{ g = \begin{bmatrix} g_1 \\ \vdots \\ g_d \end{bmatrix} \in \mathbb{R}^d \ \middle|\ g_j \in \partial(|.|)(x_j) \ \forall j = 1..d \right\}$$

$\underline{d=2}$

$$\partial\left(\|.\|_1\right)(x) = \begin{cases} [-1,1]^2 & \text{if } x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \left\{ \begin{bmatrix} 1 \\ g_2 \end{bmatrix} \middle| g_2 \in [-1,1] \right\} & \text{if } x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \text{ with } x_1 > 0 \\ \left\{ \begin{bmatrix} -1 \\ g_2 \end{bmatrix} \middle| g_2 \in [-1,1] \right\} & \text{if } x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \text{ with } x_1 < 0 \\ \vdots & \end{cases}$$

$$-\nabla \varphi_k(x^*) \in \partial(\lambda\|\cdot\|_1)(x^*)$$

$$\iff -\frac{1}{\lambda}\nabla \varphi_k(x^*) \in \partial(\|\cdot\|_1)(x^*)$$

$$h(x) \geq h(y) + g^T(y-x) \qquad g \in \partial h(y)$$

$$\frac{1}{\lambda}h(x) \geq \frac{1}{\lambda}h(y) + \left(\frac{g}{\lambda}\right)^T(y-x) \qquad \frac{g}{\lambda} \in \partial\left(\frac{1}{\lambda}h\right)(y)$$

$$-\frac{1}{\lambda}\nabla \varphi_k(x^*) \in \partial(\|\cdot\|_1)(x^*)$$

$$\iff \forall j = 1..d, \quad \left[-\frac{1}{\lambda}\nabla \varphi_k(x^*)\right]_j \in \partial(|\cdot|)(x_j^*)$$

$$\Rightarrow \quad \text{If } x_j^* = 0, \quad \left[-\frac{1}{\lambda}\nabla \varphi_k(x^*)\right]_j \in [-1,1]$$

$$\text{If } x_j^* > 0, \quad \left[-\frac{1}{\lambda}\nabla \varphi_k(x^*)\right]_j = 1$$

$$\text{If } x_j^* < 0, \quad \left[-\frac{1}{\lambda}\nabla \varphi_k(x^*)\right]_j = -1$$

$$-\frac{1}{\lambda}\nabla \varphi_k(x^*) = -\frac{1}{\lambda}\nabla f(x_k) - \frac{1}{\alpha_k\lambda}(x^* - x_k)$$

$$= \frac{1}{\alpha_k\lambda}\left(x_k - \alpha_k \nabla f(x_k) - x^*\right)$$

(i) $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k) - x^*\right]_j \in [-1,1]$ if $x_j^* = 0$

(ii) $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k) - x^*\right]_j = 1$ if $x_j^* > 0$

(iii) $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k) - x^*\right]_j = -1$ if $x_j^* < 0$

3 cases

(i) $x_j^* = 0$ and $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k)\right]_j \in [-1,1]$

(ii) $x_j^* > 0$ and $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k)\right]_j > 1$

(iii) $x_j^* < 0$ and $\frac{1}{\alpha_k \lambda}\left[x_k - \alpha_k \nabla f(x_k)\right]_j < -1$

Theorem: The solution of

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k}\|x - x_k\|_2^2 + \lambda \|x\|_1 \right\}$$

is given by $x^* \in \mathbb{R}^d$ such that

$$\forall j = 1..d, \quad [x^*]_j = \begin{cases} 0 & \text{if } [x_k - \alpha_k \nabla f(x_k)]_j \in [-\alpha_k\lambda, \alpha_k\lambda] \\ [x_k - \alpha_k \nabla f(x_k)]_j - \alpha_k\lambda & \text{if } [x_k - \alpha_k \nabla f(x_k)]_j > \alpha_k\lambda \\ [x_k - \alpha_k \nabla f(x_k)]_j + \alpha_k\lambda & \text{if } [x_k - \alpha_k \nabla f(x_k)]_j < -\alpha_k\lambda \end{cases}$$

$\longrightarrow$ Proof: Use subgradient theory on simply the formula for $\text{prox}_{\lambda\alpha_k\|\cdot\|_1}(\cdot)$

$$x_{k+1} = \text{prox}_{\lambda\alpha_k\|\cdot\|_1}\left(x_k - \alpha_k \nabla f(x_k)\right)$$

and $$\left[\text{prox}_{\lambda\|\cdot\|_1}(x)\right]_j = \begin{cases} x_j - \lambda & \text{if } x_j > \lambda \\ x_j + \lambda & \text{if } x_j < -\lambda \\ 0 & \text{if } x_j \in [-\lambda, \lambda] \end{cases}$$

<u>Note</u>: $$[x_{k+1}]_j = 0 \quad \text{if} \quad [x_k - \alpha_k \nabla f(x_k)]_j \in [-\alpha_k\lambda, \alpha_k\lambda] \quad \forall k, \forall j$$

$\lambda \longrightarrow \infty$, the condition becomes more and more likely

This update guarantees that $\|x_{k+1}\|_0 \leq \|x_k - \alpha_k \nabla f(x_k)\|_0$

A proximal gradient iteration always produces an iterate with less nonzero coordinates (or the same number) than the iterate of a gradient iteration.

ISTA: Iterative Soft-Thresholding Algorithm

→ Name for proximal gradient with ℓ_1 regularization in the signal processing community

→ Comes with theoretical convergence rates

→ Big advantage: Explicit formula for the "prox" thanks to the soft-thresholding operator

For convex problems, there is an accelerated version of ISTA called Fast ISTA, or FISTA (Beck & Teboulle, 2009)

Iteration:

$$\begin{cases} p_{k+1} = x_k + \beta_{k+1}(x_k - x_{k-1}) & \text{(Momentum step)} \\ x_{k+1} = \text{prox}_{\lambda \alpha_k \|\cdot\|_1}\left(p_{k+1} - \alpha_k \nabla f(p_{k+1})\right) \end{cases}$$

Proximal gradient steps

② Use cases of proximal gradient for ℓ_1

* LASSO (Least Absolute Selection Shrinkage Operator)

Tibshirani 1996

Linear regression + ℓ_1 regularization

$$(P_{LASSO}) \quad \underset{x \in \mathbb{R}^d}{\text{minimize}} \; \frac{1}{2}\|Ax - y\|^2 + \lambda \|x\|_1$$

(Also called basis pursuit)

$A \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\lambda > 0$

Goal of this formulation: Obtain a solution that has more zeros than that of the un-regularized problem while corresponding to a value of $\frac{1}{2}\|Ax - y\|^2$ that is good enough

⚠ (P_{LASSO}) has no closed-form solution in general

$\Rightarrow$ In practice, this problem is solved approximately using proximal gradient

* Low-rank matrix approximation

↳ If the variables form a matrix $X \in \mathbb{R}^{d_1 \times d_2}$, one can be interested in finding sparse matrices, in which case one considers problems of the form

$$\underset{X \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \quad \frac{1}{2} \| \mathcal{A}(X) - b \|^2 + \lambda \|X\|_1$$

$b \in \mathbb{R}^n$

$\mathcal{A}(X) \in \mathbb{R}^m$

$$\|X\|_1 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |X_{ij}|$$

e.g. $\mathcal{A}(X) = \left[\operatorname{trace}(A_i^T X) \right]_{i=1..m}$

↳ Another form of sparsity for matrices is low-rank structure

rank (matrix) = number of nonzero singular values

SVD: $X \in \mathbb{R}^{d_1 \times d_2}$, $\quad X = U \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_{\min(d_1,d_2)} \end{bmatrix} V^T$

$U^T U = U U^T = I_{d_1}$

$V^T V = V V^T = I_{d_2}$

$\sigma_1 \geq \ldots \geq \sigma_{\min(d_1,d_2)} \geq 0$

rank: # of nonzero σ_i's

$\mathrm{rank}(X) = \|\sigma(X)\|_0$ $\sigma(X)$: vector of singular values of X

Low-rank regularization

Use $\underbrace{\|X\|_*}_{\text{"nuclear norm"}} = \|\sigma(X)\|_1$

$$\underset{X \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \quad \frac{1}{2}\|A(X) - b\|^2 + \lambda \|X\|_*$$

$\Rightarrow$ Produce a solution with rank lower than (or equal to) that of the solution of

$$\underset{X \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \quad \frac{1}{2}\|A(X) - b\|^2$$

$\Rightarrow$ Proximal gradient applies to that setting!

* Structured sparsity regularizers (ℓ_1 aka LASSO regularization)

$\hookrightarrow$ Group LASSO / Group ℓ_1

$\forall x \in \mathbb{R}^d$, $\Omega(x) = \sum_{g \in G} \|x_g\|_2$ $\Big] \Rightarrow$ Forces sparsity of groups of variables

G is a partition of $\{1, \dots, d\}$ into groups of variables

(Ex) $x = \begin{bmatrix} x_{g_1} \\ x_{g_2} \\ x_{g_3} \end{bmatrix} \begin{matrix} \updownarrow d/3 \\ \updownarrow d/2 \\ \updownarrow d/6 \end{matrix}$ $\quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$ $\begin{matrix} g_1 = \{1,4\} \\ g_2 = \{2,3\} \\ g_3 = \{5\} \end{matrix}$

$\longrightarrow$ Group LASSO : ℓ_1 norm of $\begin{bmatrix} \|x_{g_1}\|_2 \\ \vdots \\ \|x_{g_m}\|_2 \end{bmatrix}$ $\quad |\mathcal{G}| = m$

$\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{d\}\}$: $\sum_{g \in \mathcal{G}} \|x_g\|_2 = \sum_{j=1}^{d} |x_j| = \|x\|_1$

$\mathcal{G} = \{\{1, \ldots, d\}\}$: $\sum_{g \in \mathcal{G}} \|x_g\|_2 = \|x\|_2 = \sqrt{\sum_{j=1}^{d} [x_j]^2}$

$\hookrightarrow$ Beyond that: $\quad \ell_1/\ell_\infty \quad\quad \ell_1/\ell_p \quad p \in \{1, 2, \infty\}$

- ℓ_p/ℓ_q regularization : $\sum_{g \in \mathcal{G}} \|x_g\|_\infty$, $\sum_{g \in \mathcal{G}} \|x_g\|_p$ $\quad$ ℓ_p norm

$\sum_{g \in \mathcal{G}} w_g \|x_g\|_2$ (weighted group LASSO)

- overlapping groups : problem dependent

$\Rightarrow$ All these regularizations define sparsity-inducing regularizers and different "unit balls"

$\{x \in \mathbb{R}^d \mid \Omega(x) \leq 1\}$

NB: $\Omega(x) = \frac{\lambda}{2}\|x\|_2^2 + \mu\|x\|_1$ $\quad \lambda > 0, \mu > 0$

"Elastic net regularization"

$\longrightarrow$ Can apply proximal gradient ($\text{prox}_\Omega(\cdot)$ has a closed form)

Remark: In deep learning, dropout is a way to perform sparse updates of the parameters of a neural network during training $\Rightarrow$ One form of sparse regularization

$\hookrightarrow$ Connection to coordinate descent (see next lecture)

Exercise

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \frac{\lambda}{2}\|Lx\|_2^2 \qquad f\, C^1$$

$$L = \begin{bmatrix} -2 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{bmatrix} \qquad L \in \mathbb{R}^{d \times d}, \quad L_{ij} = \begin{cases} 1 & \text{if } j = i+1 \text{ or } j = i-1 \\ -2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

a) Write down the proximal gradient iteration for that problem. Is the prox operator easy to compute?

b) Application:

$$f(x) = \frac{1}{2}\|Ax-y\|^2$$

What is the cost of solving the subproblem?

c) If $A = Id$, show that the proximal subproblem solution x_{k+1} is such that $[x_{k+1}]_j$ only depends on a subset of coordinates of $[x_k]_j$.

(<u>NB</u>: $\mathrm{prox}_{\frac{1}{2}\|L\cdot\|_2^2}(\cdot) = ?$)