Labs
Optimization for Machine Learning
Spring 2024

Saarland University
CISPA Helmholtz Center for Information Security
Sebastian Stich
TAs: Yuan Gao & Xiaowen Jiang
https://cms.cispa.saarland/optml24/

Problem Set 10 — Solutions (Proximal Methods & Compression)

1 Proximal Methods

1.1 Properties of Proximal Operator

1.1.1 Reformulation of proximal operator

Let g be proper closed convex, recall the following optimality condition:

$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{y}} g(\mathbf{y}) \Leftrightarrow 0 \in \partial g(\mathbf{x})$$

Now show that for a proper closed convex f, the proximal operator can be formulated as the following:

$$\mathbf{u} = \mathrm{prox}_f(\mathbf{x}) \Leftrightarrow \mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$$

Proof. Note that $\mathbf{u} = \mathrm{prox}_f(\mathbf{x}) \Leftrightarrow 0 \in \partial \left(\frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 + f(\mathbf{u})\right) = \mathbf{u} - \mathbf{x} + \partial f(\mathbf{u})$. Therefore $\mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$ □

1.1.2 Monotonicity of partial differential

Show that the subdifferential of a convex function $f(\mathbf{x})$ at $\mathbf{x} \in \mathbf{dom}(f)$ is a monotone operator, i.e.,

$$(\mathbf{u}-\mathbf{v})^\top(\mathbf{x}-\mathbf{y}) \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f), \mathbf{u} \in \partial f(\mathbf{x}), \mathbf{v} \in \partial f(\mathbf{y}).$$

Proof. By definition, we have

$$\begin{cases} f(y) \geq f(x) + u^T(y-x) \\ f(x) \geq f(y) + v^T(x-y) \end{cases}$$

Combining the two inequalities leads to the monotonicity. □

1.1.3 Firm nonexpansiveness

Let $f : \mathbb{R}^d \to \mathbb{R}$ be proper closed convex, write $P(\mathbf{x}) = \mathrm{prox}_f(\mathbf{x})$ and $Q(\mathbf{x}) = \mathbf{x} - P(\mathbf{x})$. Prove the following:

$$\|P(\mathbf{x}) - P(\mathbf{y})\|^2 + \|Q(\mathbf{x}) - Q(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Proof. Since $\mathbf{x} - P(\mathbf{x}) \in \partial f(P(\mathbf{x}))$ and $\mathbf{y} - P(\mathbf{y}) \in \partial f(P(\mathbf{y}))$, we apply the monotonicity of the subdifferential to get:

$$\langle \mathbf{x} - P(\mathbf{x}) - (\mathbf{y} - P(\mathbf{y})), P(\mathbf{x}) - P(\mathbf{y}) \rangle \geq 0$$

Rearranging the terms, we get

$$\langle \mathbf{x} - \mathbf{y}, P(\mathbf{x}) - P(\mathbf{y}) \rangle \geq \|P(\mathbf{x}) - P(\mathbf{y})\|^2$$

Now note that $\|P(\mathbf{x}) - P(\mathbf{y})\|^2 + \|Q(\mathbf{x}) - Q(\mathbf{y})\|^2 = \|\mathbf{x} - \mathbf{y}\| - 2\langle \mathbf{x} - \mathbf{y}, P(\mathbf{x}) - P(\mathbf{y}) \rangle + 2\|P(\mathbf{x}) - P(\mathbf{y})\|^2$. Applying the previous inequality, we get the desired result. □

1.2 LASSO

Consider the LASSO problem:

$$\min_{\mathbf{w}} \underbrace{\frac{1}{2}\|A\mathbf{w}-\mathbf{b}\|_2^2}_{f(\mathbf{w})} + \underbrace{\mu\|\mathbf{w}\|_1}_{\psi(\mathbf{w})} \tag{1}$$

Solving the LASSO problem with proximal gradient method requires the computation of the proximal operator of the following form (for some stepsize γ):

$$\mathrm{prox}_{\gamma\psi}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{z}\|^2 + \mu\|\mathbf{y}\|_1 \right\} \tag{2}$$

where $\mathbf{z} = \mathbf{w} - \gamma\nabla f(\mathbf{w}) = \mathbf{w} - \gamma A^T(A\mathbf{w}-\mathbf{b})$. Find the solution to problem (2).

Proof. By definition, we have:

$$\mathrm{prox}_{\gamma\psi}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\mu\gamma}\sum_{i\in[d]}(\mathbf{z}_i-\mathbf{y}_i)^2 + \sum_{i\in[d]}|\mathbf{y}_i| \right\}$$

We can separate this into each of the ith coordinates:

$$[\mathrm{prox}_{\gamma\psi}(\mathbf{z})]_i = \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\mu\gamma}(\mathbf{z}_i-\mathbf{y}_i)^2 + |\mathbf{y}_i| \right\}$$

The solution to this problem is given by the soft-thresholding operator:

$$S_{\mu\gamma}(\mathbf{z}_i) = \begin{cases} \mathbf{z}_i - \mu\gamma & \text{if} \quad \mathbf{z}_i > \mu\gamma \\ 0 & \text{if} \quad |\mathbf{z}_i| \le \mu\gamma \\ \mathbf{z}_i + \mu\gamma & \text{if} \quad \mathbf{z}_i < -\mu\gamma \end{cases}$$

We can verify this using Section 1.1.1. If $\mathbf{z}_i > \mu\gamma$, then $S_{\mu\gamma}(\mathbf{z}_i) = \mathbf{z}_i - \mu\gamma > 0$. Therefore $\partial(\gamma\psi)(S_{\mu\gamma}(\mathbf{z}_i)) = \mu\gamma = \mathbf{z}_i - S_{\mu\gamma}(\mathbf{z}_i)$. The other cases can be verified similarly. □

2 Compression

Recall the following definitions:

Definition 1 (ω-quantizer)**.** *We say that a (possibly randomized) mapping $\mathcal{Q}_\omega : \mathbb{R}^d \to \mathbb{R}^d$ is a quantizer if for some $\omega \ge 0$ it holds*

$$\mathbb{E}_{\mathcal{Q}_\omega}[\mathcal{Q}_\omega(\mathbf{x})] = \mathbf{x}, \quad \mathbb{E}_{\mathcal{Q}_\omega}\left[\|\mathcal{Q}_\omega(\mathbf{x})\|^2\right] \le (1+\omega)\|\mathbf{x}\|^2 \tag{3}$$

Definition 2 (δ-compressor)**.** *We say that a (possibly randomized) mapping $\mathcal{C} \colon \mathbb{R}^d \to \mathbb{R}^d$ is a contractive compression operator if for some constant $0 < \delta \le 1$ it holds*

$$\mathbb{E}\left[\|\mathcal{C}(\mathbf{x})-\mathbf{x}\|^2\right] \le (1-\delta)\|\mathbf{x}\|^2 \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{4}$$

2.1 Top-k

Show that the top-k operator is a δ-compressor. For which δ?

The $\mathrm{top}_k \colon \mathbb{R}^d \to \mathbb{R}^d$ operator is defined as

$$\left(\mathrm{top}_k(\mathbf{x})\right)_i = \begin{cases} (\mathbf{x})_{\pi(i)} & \text{if } i \le k \\ 0 & \text{otherwise} \end{cases}$$

where $k \in [d]$ is a parameter and π a permutation of the indices $\{1,\dots,d\}$, such that $(|\mathbf{x}|)_{\pi(i)} \ge (|\mathbf{x}|)_{\pi(i+1)}$ for $i = 1,\dots,d-1$. Here $(\mathbf{x})_i$ denotes the i-th coordinate of the vector $\mathbf{x}$.

Proof. It holds $\|\mathbf{x}-\mathrm{top}_k(\mathbf{x})\|^2 \leq \left(1-\frac{k}{d}\right)\|\mathbf{x}\|^2$ and the inequality is tight (consider $\mathbf{x}=\mathbf{1}$, the all-one vector). □

2.2 Rescaled Quantizer

Let $\mathcal{Q}\colon \mathbb{R}^d \to \mathbb{R}^d$ be an unbiased ω-quantizer. Show that $\frac{1}{1+\omega}\mathcal{Q}(\mathbf{x})$ is a δ-compressor. For which δ?

Proof. We observe

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{1+\omega}\mathcal{Q}(\mathbf{x})-\mathbf{x}\right\|^2 &= \mathbb{E}\left\|\frac{1}{1+\omega}\mathcal{Q}(\mathbf{x})\right\|^2 - 2\mathbb{E}\frac{1}{1+\omega}\mathcal{Q}(\mathbf{x})^\top\mathbf{x} + \|\mathbf{x}\|^2 \\
&\leq \frac{1}{1+\omega}\|\mathbf{x}\|^2 - 2\frac{1}{1+\omega}\|\mathbf{x}\|^2 + \|\mathbf{x}\|^2 \\
&= \left(1-\frac{1}{1+\omega}\right)\|\mathbf{x}\|^2 \,,
\end{aligned}
$$

where we used the two properties of a ω-quantizer (second-moment bound and unbiasedness). □