CS 2601 Linear and Convex Optimization

11. Projected gradient descent

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Outline

- Algorithm and examples

- Convergence analysis

Projected gradient descent

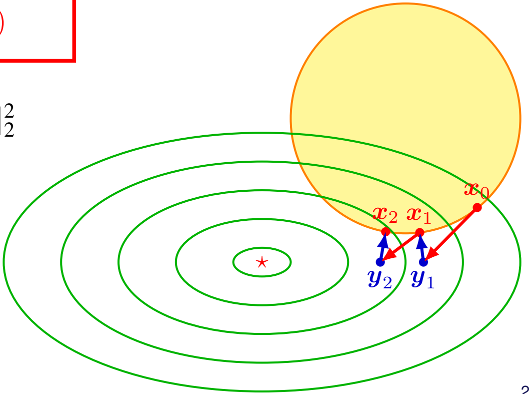
Can we apply gradient descent to a constrained problem?

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x})$$

What if $\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$ is infeasible? Project it onto X!

$$\boxed{\boldsymbol{x}_{k+1} = \mathcal{P}_X(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))}$$

where $\mathcal{P}_X(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$
is the projection of $\boldsymbol{y}$ onto X.



Useful if $\mathcal{P}_X$ can be computed efficiently.

Stopping criterion

Rewrite the projected gradient step as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\boldsymbol{g}(\boldsymbol{x}_k), \quad \text{where } \boldsymbol{g}(\boldsymbol{x}) = \frac{1}{t}(\boldsymbol{x} - \mathcal{P}_X(\boldsymbol{x} - t\nabla f(\boldsymbol{x})))$$

Lemma. For convex f and X, $\boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{0}$ iff $\boldsymbol{x}^*$ is a minimum of f over X.

Proof. Recall from slide 30 of §3, $\hat{\boldsymbol{x}} = \mathcal{P}_X(\boldsymbol{x})$ iff

$$\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \boldsymbol{z} - \hat{\boldsymbol{x}} \rangle \le 0, \quad \forall \boldsymbol{z} \in X$$

so

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{0} &\iff \boldsymbol{x}^* = \mathcal{P}_X(\boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*)) \\ &\iff \langle \boldsymbol{x}^* - t\nabla f(\boldsymbol{x}^*) - \boldsymbol{x}^*, \boldsymbol{z} - \boldsymbol{x}^* \rangle \le 0, \quad \forall \boldsymbol{z} \in X \\ &\iff \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{z} - \boldsymbol{x}^* \rangle \ge 0, \quad \forall \boldsymbol{z} \in X \\ &\iff \boldsymbol{x}^* \text{ is a minimum of } f \text{ over } X \end{aligned}$$

Note. $\boldsymbol{g}(\boldsymbol{x})$ plays a similar role as $\nabla f(\boldsymbol{x})$ does in gradient descent. We can stop when $\boldsymbol{g}(\boldsymbol{x}_k)$ is small, or equivalently when $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ is small.

Examples

$\mathcal{P}_X$ can be efficiently computed for the following constraints.

- ℓ_2 constraint

$$\|\boldsymbol{x}\|_2 \leq t$$

  e.g. ridge regression

- box constraint

$$\boldsymbol{a} \leq \boldsymbol{x} \leq \boldsymbol{b} \quad i.e. \quad a_i \leq x_i \leq b_i, i = 1, 2, \ldots, n$$

  Special case. ℓ_∞ constraint $\|\boldsymbol{x}\|_\infty \leq t$.

- affine constraint 这个很重要

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

- ℓ_1 constraint

$$\|\boldsymbol{x}\|_1 \leq t$$

  e.g. Lasso

Projection onto ℓ_2 ball

For $X = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \le t\}$,

$$\mathcal{P}_X(\boldsymbol{y}) = \min\left\{1, \frac{t}{\|\boldsymbol{y}\|_2}\right\}\boldsymbol{y}$$

Proof. Solve

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{x}\|_2^2 \le t^2 \end{aligned}$$

By the KKT conditions, there exists $\mu \ge 0$ s.t.

$$\boldsymbol{x} - \boldsymbol{y} + 2\mu\boldsymbol{x} = 0 \implies \boldsymbol{x} = \frac{\boldsymbol{y}}{1 + 2\mu} \propto \boldsymbol{y}$$

- If $\|\boldsymbol{y}\|_2 \le t$, then $\mu = 0$ and $\boldsymbol{x} = \boldsymbol{y}$.
- If $\|\boldsymbol{y}\|_2 > t$, then $\mu > 0$ and $\boldsymbol{x} = \frac{t}{\|\boldsymbol{y}\|}\boldsymbol{y}$.

Projection onto box

For $X = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a} \le \boldsymbol{x} \le \boldsymbol{b}\}$,

$$\mathcal{P}_X(\boldsymbol{y}) = \min\{\boldsymbol{b}, \max\{\boldsymbol{a}, \boldsymbol{y}\}\}$$

i.e.

$$x_i = \begin{cases} a_i, & \text{if } y_i \le a_i \\ y_i, & \text{if } a_i \le y_i \le b_i \\ b_i, & \text{if } y_i \ge b_i \end{cases}$$

Note. Each component is projected independently.

Proof. The problem is decomposable,

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ \text{s.t.} & \boldsymbol{a} \le \boldsymbol{x} \le \boldsymbol{b} \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \min\limits_{x_i} & \frac{1}{2}(x_i - y_i)^2 \\ \text{s.t.} & a_i \le x_i \le b_i \end{array}, \quad i = 1, 2, \ldots, n$$

Projection onto ℓ_1 ball hard

For $X = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le t\}$, need to solve

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = \frac{1}{2}\sum_{i=1}^{n}(x_i - y_i)^2 \\ \text{s.t.} \quad & \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n}|x_i| \le t \end{aligned} \qquad (\dagger)$$

不可导，不好直接用KKT

$\mathcal{P}_X$ has no closed-form solution, but can be computed efficiently.

Observation 1. By symmetry, the general case can be reduced to the case $\boldsymbol{y} \ge \mathbf{0}$ by

$$\boldsymbol{x} = \operatorname{sgn}(\boldsymbol{y})\,\mathcal{P}_X(\operatorname{abs}(\boldsymbol{y}))$$

e.g. if $(x_1, x_2) = \mathcal{P}_X(y_1, y_2)$, then $(-x_1, x_2) = \mathcal{P}_X(-y_1, y_2)$.

Observation 2. If $\boldsymbol{y} \ge \mathbf{0}$, then the solution $\boldsymbol{x} \ge \mathbf{0}$. If $x_i < 0$ for some i, then replacing x_i by $-x_i$ yields a better solution.

Projection onto ℓ_1 ball (cont'd)

Now focus on the case $\boldsymbol{y} \geq \boldsymbol{0}$. The problem reduces to

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\sum_{i=1}^{n}(x_i - y_i)^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i \leq t \\ & x_i \geq 0, \quad i = 1, 2, \ldots, n \end{aligned}$$

By the KKT conditions, there exists $\mu_i \geq 0$, $i = 0, 1, \ldots, n$ s.t.

$$x_i - y_i + \mu_0 - \mu_i = 0, \quad i = 1, 2, \ldots, n$$

Thus (cf. slide 18 of §10)

$$x_i = (y_i - \mu_0)^+$$

which is soft-thresholding with unknown μ_0.

Projection onto ℓ_1 ball (cont'd)

μ_0 is determined by the constraint

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} (y_i - \mu_0)^+ \leq t$$

- If $\|\boldsymbol{y}\|_1 \leq t$, then $\mu_0 = 0$, $\boldsymbol{x} = \boldsymbol{y}$.
- If $\|\boldsymbol{y}\|_1 > t$, then $\mu_0 > 0$ and $\boldsymbol{x}$ can be found in a similar way as in the water filling solution.
  - Sort $\boldsymbol{y}$ s.t. $y_1 \geq y_2 \geq \cdots \geq y_n$
  - Let

    $$c_k = \frac{1}{k}\left(\sum_{i=1}^{k} y_i - t\right)$$

    Then

    $$\mu_0 = c_{k_0}$$

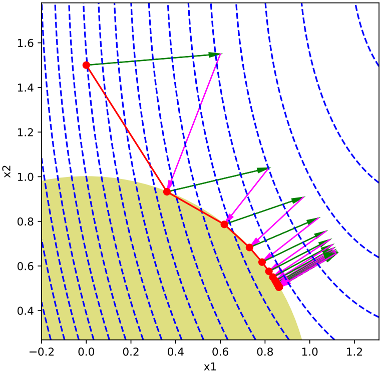
    where

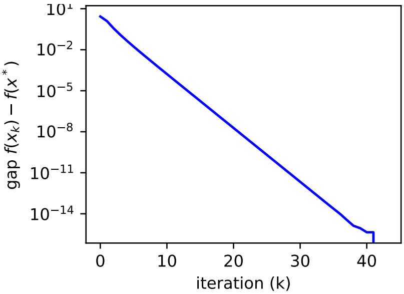
    $$k_0 = \max\{k : c_k \leq y_k\}$$

Example: Ridge regression

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{w}\|_2 \le t \end{aligned} \qquad \boldsymbol{X} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}, \quad t = 1$$

Step size $t = 0.1$, $\boldsymbol{w}_0 = (0, 1.5)^T$, $\boldsymbol{w}^* \approx (0.86270563, 0.50570644)^T$.



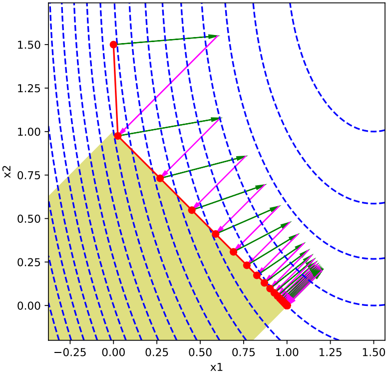


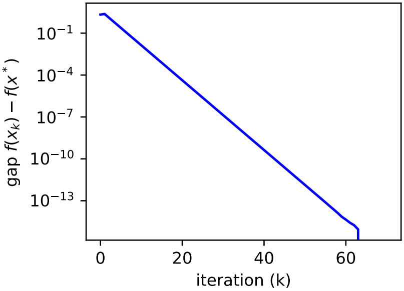
Example: Lasso

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 \qquad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \le t$$

$$\boldsymbol{X} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}, \quad t = 1$$

Step size $t = 0.1$, $\boldsymbol{w}_0 = (0, 1.5)^T$, $\boldsymbol{w}^* = (1, 0)^T$.





Outline

- Algorithm and examples
- Convergence analysis

Convergence analysis

Theorem. Let X be a nonempty convex set, and f an L-smooth and m-strongly convex[1] function over X. Let $\boldsymbol{x}^*$ be a minimum of f over X. The sequence $\{\boldsymbol{x}_k\}$ produced by projected gradient descent with constant step size $t = \frac{1}{L}$ has the following properties.

1. $f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k)$ and

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \le \frac{L}{2k}\|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2$$

2. $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 \le (1 - \frac{m}{L})\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2$, and hence

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \le (1 - \frac{m}{L})^k\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$$

Note. The results are similar to those on slide 13 of §8. In fact, we will see that projected gradient descent can be considered as a special case of proximal gradient descent.

[1] we allow $m = 0$ for general convex f.

Connection to proximal gradient descent

Define the indicator I_X of a set X by

$$I_X(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in X \\ +\infty, & \text{if } \boldsymbol{x} \notin X \end{cases}$$

Note. I_X is a convex function iff X is a convex set.

The proximal operator for I_X is simply the projection onto X,

$$\begin{aligned} \operatorname{prox}_{I_X}(\boldsymbol{y}) &= \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + I_X(\boldsymbol{x}) \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{x}\in X} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \\ &= \mathcal{P}_X(\boldsymbol{y}) \end{aligned}$$

Connection to proximal gradient descent (cont'd)

Note

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}) \iff \min_{\boldsymbol{x}} \{f(\boldsymbol{x}) + I_X(\boldsymbol{x})\}$$

Since $I_X(\boldsymbol{x}) = t_k I_X(\boldsymbol{x})$ for $t_k > 0$,

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \mathcal{P}_X(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \\
&= \mathrm{prox}_{I_X}(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \\
&= \mathrm{prox}_{t_k I_X}(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))
\end{aligned}$$

Projected gradient descent for $\min_{\boldsymbol{x} \in X} f(\boldsymbol{x})$ is the same as proximal gradient descent for $\min_{\boldsymbol{x}} \{f(\boldsymbol{x}) + I_X(\boldsymbol{x})\}$!

By restricting to $\boldsymbol{x} \in X$, the convergence analysis for proximal gradient descent applies to projected gradient descent without further change.