Convex Optimization

Part 2: Subgradient method (2/2)

Namhoon Lee

POSTECH

28 Sep 2022

Admin

Assignment 1 is being graded.

- The result will be uploaded on PLMS.
- You can check your result with TAs until Assignment 2 is due.

Assignment 2 is out already.

- Due by Wed 19 Oct.

No class on Mon 10 Oct.

Be reminded of pop quizzes.

Least squares with l_1-regularization

Given some data $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^n$ and linear prediction model $\hat{y} = \beta^\top x$, consider the least squares with l_1-regularization

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where λ is the regularization coefficient.

- If λ is sufficiently large, the solution can be sparse (useful for feature selection).
- Why? How does it compare to l_2-regularization?
- Optimality condition?

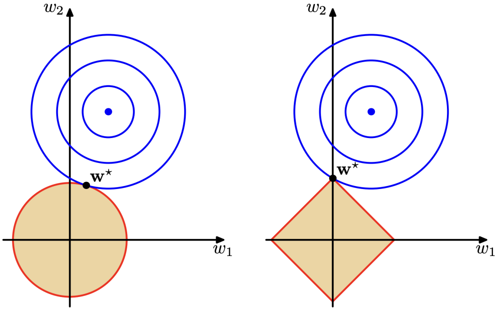


Figure: l_2 vs l_1 regularization. Figure taken from Bishop.

Apply the optimality condition to both cases

For l_2-regularization

$$0 = -X_i^\top (y - X\beta) + \lambda\beta_i$$

- It is unlikely to be satisfied for $\beta_i = 0$.

For l_1-regularization

$$0 \in -X_i^\top (y - X\beta) + \lambda[-1, 1]$$

- The chance is better now since $|X_i^\top (y - X\beta)| \in \lambda$ is more likely (with large λ).

To complete, the optimality condition for lasso

$$
\begin{aligned}
0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) \quad &\iff \quad 0 \in -X^\top(y - X\beta) + \lambda\partial\|\beta\|_1 \\
&\iff \quad \beta = (X^\top X)^{-1}(X^\top y - \lambda z)
\end{aligned}
$$

where $z = \partial\|\beta\|_1$, *i.e.*,

$$
z_i = \begin{cases} \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ \in [-1, 1] & \text{if } \beta_i = 0 \end{cases}
$$

- This does not provide the optimal solution; rather it is a characterization; but still it could be useful.

Subgradient method

Consider minimizing f that is convex but not necessarily differentiable.

Subgradient method

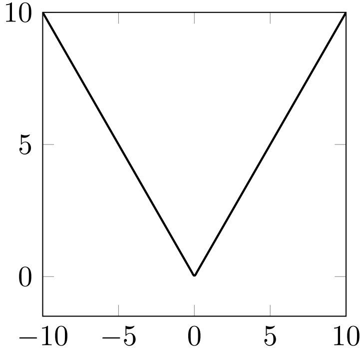
Start with some initial point x_1, repeat the following update step iteratively

$$x_{t+1} = x_t - \eta_t g_t$$

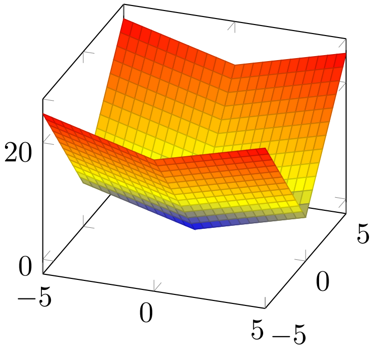
and stop at some point. Here $g_t \in \partial f(x_t)$, *i.e.*, any subgradient of f at x_t.

- One can keep $x_{t,\text{best}}$ instead of x_T because subgradient method is not necessarily a descent method (hence the name); *i.e.*, it can increase the objective (why?).

$$f(x) = |x|$$



$$f(x_1, x_2) = |x_1| + 4|x_2|$$



Step size

Fixed step size

- $\eta_t = \bar{\eta}$ for all $t = 1, 2, \ldots$

Diminishing step size

- $\eta_t \to 0$ as $t \to \infty$; specifically η_t under the following conditions

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

  i.e., η_t decreases to 0 but not too fast.

Optimal step size

- $\eta_t = (f(x_t) - f^*)/\|g_t\|_2^2$

Convergence analysis

Theorem

For f convex and Lipschitz continuous with parameter $G > 0$ (or bounded subgradient), subgradient method with step size η satisfies

$$f(x_{t,\textsf{best}}) - f^* \le \frac{R^2}{2\eta T} + \frac{G^2\eta}{2}$$

where $R = \|x_1 - x^*\|_2$.

- For fixed step size it fails to converge to 0 error (*i.e.*, $G^2\eta/2$-suboptimal).
- For diminishing step size $\eta \propto 1/\sqrt{t}$, we can get $\mathcal{O}(1/\sqrt{T})$ convergence rate which is slower than gradient descent (it makes sense why? smaller step sizes).
  - It does not accelerate even if we add momentum (later).

We prove for a general case in which subgradient method runs with step size η_t that decreases to 0 as t increases; *i.e.*, as $\eta_t \to 0$ as $t \to \infty$ and further $\sum_{t=1}^{\infty} \eta_t = \infty$.

Proof.

Following the similar convergence proof for gradient descent and using the bounded gradient assumption we arrive

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\eta_t(f(x_t) - f(x^*)) + \eta_t^2 G^2$$

Summing for T iterations, lower-bounding $f(x_t)$ with $f(x_{t,\text{best}})$, and rearranging terms

$$f(x_{t,\text{best}}) - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t}$$

For $\eta_t \propto 1/\sqrt{t}$, $\sum_{t=1}^{T} \eta_t^2 / \sum_{t=1}^{T} \eta_t \to 0$, indicating that $f(x_{t,\text{best}})$ converges to f^*. □

Polyak step size (Polyak 1987)

If f^* is known, one can come up with optimal step sizes

$$\eta_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$$

which is obtained by minimizing the intermediate result of progress in one iteration from the proof.

Applying this step size will give

$$f(x_{t,\text{best}}) - f^* \le \frac{RG}{\sqrt{T}}$$

which achives the optimal result; the convergence rate is still $\mathcal{O}(1/\sqrt{T})$.

- A simple variant can get near optimal rates without knowledge of f^* (Hazan and Kakade 2019).

On the subgradient method

While subgradient method can be applied nearly all non-smooth convex optimization, it is very slow ($\mathcal{O}(1/\sqrt{t})$).

- This is an optimal rate, and it does not improve with a momentum scheme.

Instead we could make use of some structure of the problem, which could give us better convergence rates (next time).

Any questions?