ℓ_1 Norm Projection

Kaiwen Wu
School of Computer Science, University of Waterloo
kaiwen.wu@uwaterloo.ca

May 2020

Abstract

We present a simple derivation of ℓ_1 norm projection. The note is intended for self-reference and does not involve messy calculation on KKT conditions.

1 Introduction

We consider the following ℓ_1 norm projection problem:

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \;\; \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 \\ &\text{subject to} \;\; \|\mathbf{x}\|_1 \le 1. \end{aligned}$$

This problem is known to be solved in $O(n \log n)$ time complexity. A highly cited reference nowadays is written by Duchi et al. (2008), though an approach based on the same idea dates back to the 1970s (Held et al. 1974).

2 Method

Write down the dual problem via the method of Lagrange multipliers:

$$\begin{aligned} &\text{maximize} \;\; g(\lambda) \\ &\text{subject to} \;\; \lambda \ge 0, \end{aligned}$$

where

$$g(\lambda) = \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \lambda\left(\|\mathbf{x}\|_1 - 1\right).$$

Evaluating the $g(\lambda)$ is equivalent to computing a proximal operator $\operatorname{prox}_{\lambda\|\cdot\|_1}(\mathbf{z})$, *i.e.*, the soft thresholding operator. By KKT conditions, primal and dual solutions satisfy

$$\begin{aligned} \mathbf{x}^\star &= \operatorname{prox}_{\lambda^\star\|\cdot\|_1}(\mathbf{z}) \\ &= \operatorname{sign}(\mathbf{z})\left(|\mathbf{z}| - \lambda^\star\right)_+. \end{aligned}$$

So the problem is reduced to computing $\lambda^\star$. Thanks to the uniqueness of the proximal operator (and Danskin's theorem), $g(\lambda)$ is differentiable and the derivative is

$$\begin{aligned} g'(\lambda) &= \|\operatorname{prox}_{\lambda\|\cdot\|_1}(\mathbf{z})\|_1 - 1 \\ &= \sum_{i=1}^{n}\left(|z_i| - \lambda\right)_+ - 1. \end{aligned}$$

Since $g(\lambda)$ is concave, any stationary point is optimal (though the reverse is not true). Notice that $g'(\lambda)$ is a univariate non-increasing function, thus we can use bisection method to find the stationary point very efficiently (and it still works when the optimum is achieved at the boundary). In particular, the following proposition gives the starting interval of bisection method.

Proposition 1 *Dual solution satisfies* $0 \leq \lambda^\star \leq \|\mathbf{z}\|_\infty$.

Proof: $\lambda^\star \geq 0$ is the dual feasible constraint. $\lambda^\star \leq \|\mathbf{z}\|_\infty$ because otherwise $g'(\lambda^\star) < 0$. The dual objective can be strictly improved. ■

Although the bisection method is simple, the message from it is interesting: whenever the proximal operator of a convex function is easy to compute, the projection operator is also easy to compute.

2.1 Compute $\lambda^\star$ Directly

WLOG, we assume that entries of $\mathbf{z}$ are nonnegative and sorted, *i.e.*, $z_1 \geq z_2 \geq \cdots \geq z_n \geq 0$, as the dual solution is invariant to the sign and the order of z_i.

Proposition 2 *Define* $\mathcal{I} = \left\{ j : \sum_{i=1}^{j} (z_i - z_j) < 1 \right\}$ *and* $k = |\mathcal{I}|$. *We have*

$$\lambda^\star = \frac{1}{k} \left(\sum_{i=1}^{k} z_i - 1 \right)_+ . \tag{1}$$

The above expression is well-defined as $1 \in \mathcal{I}$ *is always nonempty.*

Proof: Note that $\sum_{i=1}^{j}(z_i - z_j)$ is non-decreasing w.r.t. j. Indices in $\mathcal{I}$ are consecutive, *i.e.*, $\mathcal{I} = \{1, 2, \cdots k\}$. In particular, if $k < n$, then $\sum_{i=1}^{k+1}(z_i - z_{k+1}) \geq 1$. Thus $g'(z_k) < 0$ and $g'(z_{k+1}) \geq 0$. There is a stationary point between z_k and z_{k+1}. Simple calculation gives $\lambda^\star = \frac{1}{k}\left(\sum_{i=1}^{k} z_i - 1\right)$.
If $k = n$, either there is a stationary point between z_n and 0 or there is no stationary point, in which case $\lambda^\star = 0$. In both cases the expression (1) works out. ■

Remark 1 *Since the expression* (1) *depends on the cardinality of* $\mathcal{I}$, *one might wonder whether it is numerically stable when there exists some* j *such that* $\sum_{i=1}^{j}(z_i - z_j) \approx 1$. *It turns out that the cardinality* k *is not continuous w.r.t.* $\mathbf{z}$. *But* $\lambda^\star$ *is still continuous w.r.t.* $\mathbf{z}$. *Consider the case where* $\sum_{i=1}^{k+1}(z_i - z_{k+1}) \approx 1$. *We have* $\lambda^\star = \frac{1}{k}\left(\sum_{i=1}^{k} z_i - 1\right)_+ = \frac{1}{k}\left(\sum_{i=1}^{k}(z_i - z_{k+1}) + kz_{k+1} - 1\right)_+ \approx z_{k+1}$, *which is continuous in* $\mathbf{z}$.

References

[1] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. "Efficient projections onto the ℓ_1-ball for learning in high dimensions". In: *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*. 2008, pp. 272–279.

[2] Michael Held, Philip Wolfe, and Harlan P. Crowder. "Validation of subgradient optimization". In: *Mathematical Programming* 6 (1974), pp. 62–88.