# $\ell_1$ Norm Projection

Kaiwen Wu[*]

May 2020[†]

**Abstract**

$\ell_1$ norm projection arises frequently from sparsity, and is often a key building block of more complex algorithms. In this note, we present the $\ell_1$ norm projection algorithm in Duchi et al. (2008). Unlike common derivations, this note aims to present it in a clean way that avoids messy calculations of the KKT conditions.

## 1 Introduction

We consider the following $\ell_1$ norm projection problem:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|^2 \quad \text{subject to } \|\mathbf{x}\|_1 \le 1,$$

where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_1$ is the $\ell_1$ norm. Namely, we aim to find a point in a $n$-dimensional unit $\ell_1$ norm ball that is closest to a given point $\mathbf{z}$. This $\ell_1$ norm projection often arises from sparsity, e.g., compressed sensing.

In principle, the $\ell_1$ norm projection problem can be solved by any convex optimization methods. However, a generic convex optimization solver is not very efficient. A standard interior-point method, e.g., a path-following method with inner-loops solved by Newton's method, takes $\tilde{\mathcal{O}}(n^3)$ time for a $n$-dimensional problem. However, there exists a faster algorithm running in $O(n \log n)$ time and computing the exact optimum. A highly cited reference nowadays is by Duchi et al. (2008), although an approach based on the same idea dates back to the 1970s (Held et al., 1974).

## 2 Reduction to a Univariate Dual

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

$$\text{maximize } g(\lambda) \quad \text{subject to } \lambda \ge 0,$$

where $g(\lambda) = \min_{\mathbf{x}} L(\mathbf{x}, \lambda)$ is the minimum of the Lagrangian over $\mathbf{x}$, and the Lagrangian writes

$$L(\mathbf{x}, \lambda) = \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|^2 + \lambda(\|\mathbf{x}\|_1 - 1).$$

[*] David R. Cheriton School of Computer Science, University of Waterloo. Email: kaiwen.wu@uwaterloo.ca

[†] This version is polished December 2024.

Minimizing the Lagrangian, or equivalently evaluating the dual objective value $g(\lambda)$, is reduced to a proximal operator (the soft thresholding operator in this case):

$$\operatorname*{argmin}_{\mathbf{x}} L(\mathbf{x}, \lambda) = \operatorname{prox}_{\lambda\|\cdot\|_1}(\mathbf{z}) = \operatorname{sign}(\mathbf{z})(|\mathbf{z}| - \lambda)_+,$$

where $(\,\cdot\,)_+ = \max\{0, \cdot\}$ is applied element-wise. By the KKT conditions, the primal solution $\mathbf{x}^*$ and the dual solution $\lambda^*$ satisfy

$$\mathbf{x}^* = \operatorname{sign}(\mathbf{z})(|\mathbf{z}| - \lambda^*)_+.$$

Hence, the primal solution can be recovered uniquely from the dual solution $\lambda^*$. A critical observation is that the dual objective $g(\lambda)$ is differentiable thanks to Danskin's theorem.[1] Moreover, Danskin's theorem asserts that the derivative is of the dual objective is

$$g'(\lambda) = \|\operatorname{sign}(\mathbf{z})(|\mathbf{z}| - \lambda)_+\|_1 - 1 = \sum_{i=1}^{n} \left(|z_i| - \lambda\right)_+ - 1, \tag{1}$$

which is a piece-wise linear function. As a result, the dual objective is not twice-differentiable.

The objective $g(\lambda)$ is concave and its stationary point (if exists) is exactly the dual solution $\lambda^*$. Note that the derivative $g'(\lambda)$ is monotonically decreasing. Thus, we can use the bisection method to search the stationary point efficiently.[2] An edge case might occur when no stationary point exists in the domain $\lambda \geq 0$, in which case $\lambda^* = 0$. But this edge case can be handled easily by a careful implementation. To start the bisection method, we need an interval containing the dual solution, which is given as follows.

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

*Proof.* The left inequality $\lambda^* \geq 0$ is trivial since it is the domain of the dual problem. For the right inequality $\lambda^* \leq \|\mathbf{z}\|_\infty$, assume on the contrary that $\lambda^* > \|\mathbf{z}\|_\infty = \max_{1\leq i\leq n}|z_i|$. Then, we have $g'(\lambda^*) = -1$ by (1). The derivative at $\lambda^*$ is negative and thus the dual objective value can be strictly improved. Contradiction. □

## 3 Solve the Dual in a Closed-Form

The bisection method finds an $\epsilon$-approximate dual solution in $\mathcal{O}\big(n \log \frac{1}{\epsilon}\big)$ time. However, there exist methods that compute the exact dual solution in $\mathcal{O}(n \log n)$ time. Without loss of generality, we assume that all entries of $\mathbf{z}$ are nonnegative and sorted in descending order, i.e., $0 \leq z_n \leq z_{n-1} \leq \cdots \leq z_1$, as the dual solution $\lambda^*$ is invariant to the signs and the order of $z_i$'s.

**Proposition 2.** *Let* $k = \max\big\{i \in [n] : \sum_{j=1}^{i}(z_j - z_i) < 1\big\}$. *Then, the dual solution* $\lambda^*$ *is of the form*

$$\lambda^* = \frac{1}{k}\left(\sum_{i=1}^{k} z_i - 1\right)_+, \tag{2}$$

*where we recall that* $(\,\cdot\,)_+ = \max\{0, \cdot\}$.

*Proof.* Evaluate the derivative $g'(\lambda)$ at $\lambda = z_i$. We obtain

$$g'(z_i) = \underbrace{(z_1 - z_i) + (z_2 - z_i) + \cdots + (z_i - z_i)}_{i} + \underbrace{0 + \cdots + 0}_{n-i} - 1 = \sum_{j=1}^{i}(z_j - z_i) - 1.$$

[1] One can verify this claim by manually calculating the dual objective explicitly, albeit tedious.

[2] Another implication is that the projection operator is not much harder to compute than the proximal operator. Because the projection operator can be simulated by the proximal operator with an extra logarithmic factor in the running time.

Thus, $z_k$ is the smallest break point among $z_n \leq z_{n-1} \leq \cdots \leq z_1$ whose derivative is negative. We discuss two cases based on the value of $g'(0) = \sum_{j=1}^{n} z_j - 1$.

**Case 1:** $g'(0) \leq 0$. Since $g'(\lambda)$ is monotonically decreasing, we have $g'(\lambda) \leq 0$ for all $\lambda \geq 0$. Hence, $g(\lambda)$ is decreasing as well, which implies its maximizer is $\lambda^* = 0$. Once can verify that (2) checks out.

**Case 2:** $g'(0) > 0$. In this case, $g'(\lambda)$ has a unique stationary point in the domain $\lambda \geq 0$. Recall that $z_k$ is the smallest break point whose derivative is negative. Namely, we have

$$g'(0) \geq g'(z_n) \geq \cdots > g'(z_{k+1}) \geq 0 > g'(z_k) \geq \cdots \geq g(z_2) \geq g(z_1).$$

Thus, the unique stationary point falls into the interval $[z_{k+1}, z_k]$ (for convenience define $z_{n+1} = 0$ to handle the edge case $k = n$). Inside this interval, the derivative writes

$$g'(\lambda) = -1 - k \cdot \lambda + \sum_{j=1}^{k} z_j, \quad \lambda \in [z_{k+1}, z_k].$$

Solve the equation $g'(\lambda) = 0$. The root is $\lambda^* = \frac{1}{k}\left(\sum_{j=1}^{k} z_j - 1\right)$. Again, Equation (2) checks out. □

# 4 Discussion

The main reason, from the perspective of this note, why $\ell_1$ norm projection admits efficient algorithms is because its dual is univariate—optimizing univariate objectives is typically easy.

The idea of formulating univariate dual problems is powerful and has been applied to other settings. A generic method for $\ell_p$ norm projection (where $p > 1$) is developed by applying Newton's method to a twice-differentiable univariate dual problem (Won et al., 2023). The fastest algorithm to date for computing the proximal operator of the matrix perspective function is based on exactly the same idea (Won, 2020). In addition, I personally have written a paper on projection to Wasserstein balls, where the projection is solved by the optimizing carefully crafted univariate dual problems (Wu et al., 2020).

# References

Held, M., Wolfe, P., & Crowder, H. P. (1974). Validation of subgradient optimization. *Mathematical Programming*, *6*, 62–88 (page 1).

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the $\ell_1$-ball for learning in high dimensions, In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*. (Page 1).

Won, J.-H. (2020). Proximity Operator of the Matrix Perspective Function and its Applications, In *Advances in Neural Information Processing Systems*, Curran Associates, Inc. (Page 3).

Wu, K., Wang, A., & Yu, Y. (2020). Stronger and Faster Wasserstein Adversarial Attacks, In *Proceedings of the 37th International Conference on Machine Learning*, PMLR. (Page 3).

Won, J.-H., Lange, K., & Xu, J. (2023). A unified analysis of convex and non-convex $\ell_p$-ball projection problems. *Optimization letters*, *17*(5), 1133–1159 (page 3).