This is the problem I'm trying to solve:

$$f(\boldsymbol{z}) = \frac{a}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + \frac{b}{2}\|\boldsymbol{P}\boldsymbol{z} - \boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{1}$$

$$\arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) \tag{2}$$

where $\boldsymbol{P}$ is a nonlinear operator, specifically a quantized projection operator. That is, $\boldsymbol{P}(\cdot) = \boldsymbol{W}^T\,\text{Quantize}(\boldsymbol{W}\cdot)$, where $\boldsymbol{W}^T\boldsymbol{W}$ is a projection operator.

$$\text{Quantize}(\boldsymbol{y}) = \text{round}(\frac{\boldsymbol{y}}{\boldsymbol{q}}) * \boldsymbol{q} \tag{3}$$

(Division here is element-by-element).

In an early attempt to solve my problem, I pretended $\boldsymbol{P}$ was a linear projection operator and came up with an approximate solution to my original problem.

$$\boldsymbol{z}_{\text{approx}} = \boldsymbol{x} + \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) \tag{4}$$

Now, plugging $\boldsymbol{z} = \boldsymbol{z}_{\text{approx}} + \Delta\boldsymbol{z}$ back into the original equation, I have the expression:

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|\boldsymbol{P}(\boldsymbol{x}+\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z})-\boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{5}$$

In analyzing this expression, I have found it helpful to define a function $\boldsymbol{\epsilon}(\Delta\boldsymbol{z})$:

$$\boldsymbol{\epsilon}(\Delta\boldsymbol{z}) = \boldsymbol{P}(\boldsymbol{x} + \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \Delta\boldsymbol{z}) - \boldsymbol{P}\boldsymbol{x} - \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) \tag{6}$$

$$\boldsymbol{P}(\boldsymbol{x} + \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \Delta\boldsymbol{z}) = \boldsymbol{P}\boldsymbol{x} - \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \boldsymbol{\epsilon}(\Delta\boldsymbol{z}) \tag{7}$$

Returning to the objective function:

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|\boldsymbol{P}\boldsymbol{x}-\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})-\boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{8}$$

Simplifying

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|-\frac{a+b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})\|_2^2 \tag{9}$$

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|-\frac{a}{a+b}(\boldsymbol{P}\boldsymbol{s}-\boldsymbol{P}\boldsymbol{x})+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})\|_2^2 \tag{10}$$

I have 2 objective terms, and it will help to give them names:

$$f_1(\Delta \boldsymbol{z}) = \frac{a}{2} \| \frac{b}{a+b} (\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \Delta \boldsymbol{z} \|_2^2 \tag{11}$$

$$f_2(\Delta \boldsymbol{z}) = \frac{b}{2} \| - \frac{a}{a+b} (\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \boldsymbol{\epsilon}(\Delta \boldsymbol{z}) \|_2^2 \tag{12}$$

$$f(\boldsymbol{z}) = f_1(\boldsymbol{z} - \boldsymbol{z}_{\text{approx}}) + f_2(\boldsymbol{z} - \boldsymbol{z}_{\text{approx}}) \tag{13}$$

Now for some observations:

1. Adding a component to $\Delta \boldsymbol{z}$ that is orthogonal to the span of the columns of $\boldsymbol{W}^T$ increases the first term of the objective f_1 without affecting the second term f_2.

$$f_1(\Delta \boldsymbol{z}) \geq f_1(\boldsymbol{W}^T \boldsymbol{W} \Delta \boldsymbol{z}) \tag{14}$$

$$f_2(\Delta \boldsymbol{z}) = f_2(\boldsymbol{W}^T \boldsymbol{W} \Delta \boldsymbol{z}) \tag{15}$$

   Therefore,

$$(\mathbf{I} - \boldsymbol{W}^T \boldsymbol{W})(\Delta \boldsymbol{z})_{\text{optimal}} = 0 \tag{16}$$

2. In the simplified case of no quantization $\boldsymbol{P} = \boldsymbol{W}^T \boldsymbol{W}$:

$$\boldsymbol{\epsilon}(\Delta \boldsymbol{z}) = \boldsymbol{W}^T \boldsymbol{W} \Delta \boldsymbol{z} \tag{17}$$

   And so, if the quantization process is removed,

$$(\Delta \boldsymbol{z})_{\text{optimal}} = 0 \tag{18}$$

3. For $\alpha \in [0, 1]$:

$$f_2(\alpha \boldsymbol{\epsilon}(0)) = f_2(0) \tag{19}$$

   Furthermore, this is true even if I scale elements of $\boldsymbol{W}\boldsymbol{\epsilon}(0)$ individually: For $\alpha_i \in [0, 1]$:

$$f_2(\boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\alpha}) \boldsymbol{W} \boldsymbol{\epsilon}(0)) = f_2(0) \tag{20}$$

4. There are certain choices for $\boldsymbol{\alpha}$ from the previous observation that will decrease the first objective term f_1:

$$\alpha_i = \begin{cases} 1 & \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W} \boldsymbol{\epsilon}(0)) = -\operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x})) \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

   Using the $\boldsymbol{\alpha}$ defined above:

$$f_1(\boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\alpha}) \boldsymbol{W} \boldsymbol{\epsilon}(0)) \leq f_1(0) \tag{22}$$

5. The last couple of operations have focused on decreasing f_1 without affecting f_2. Here, I observe it is also possible to select a $\Delta \boldsymbol{z}$ that deceases f_2 by more than it decreases f_1.

$$\beta_i = \begin{cases} 1+\nu & \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W} \boldsymbol{\epsilon}(\boldsymbol{q}/2)) = \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x})) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where ν is an arbitrarily small number to ensure the rounding occurs in the proper direction.

Using the $\boldsymbol{\beta}$ defined above:

$$f_1(\boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\beta}) \boldsymbol{W} \boldsymbol{\epsilon}(\frac{\boldsymbol{W}^T \boldsymbol{q}}{2})) - f_1(0) \leq f_2(0) - f_2(\boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\beta}) \boldsymbol{W} \boldsymbol{\epsilon}(\frac{\boldsymbol{W}^T \boldsymbol{q}}{2})) \tag{24}$$

6. Finally, the last couple observations can be combined for the optimal solution:

$$(\Delta \boldsymbol{z})_{\text{optimal}} = \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\beta}) \boldsymbol{W} \boldsymbol{\epsilon}(\frac{\boldsymbol{W}^T \boldsymbol{q}}{2}) + \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\alpha}) \boldsymbol{W} \boldsymbol{\epsilon}(0) \tag{25}$$

where

$$\alpha_i = \begin{cases} 1 & \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W} \boldsymbol{\epsilon}(0)) = -\operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x})) \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$\beta_i = \begin{cases} 1+\nu & \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W} \boldsymbol{\epsilon}(\boldsymbol{q}/2)) = \operatorname{sign}(\boldsymbol{e}_i^T \boldsymbol{W}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x})) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

$$\boldsymbol{z}_{\text{optimal}} = \boldsymbol{x} + \frac{b}{a+b}(\boldsymbol{P}\boldsymbol{s} - \boldsymbol{P}\boldsymbol{x}) + \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\beta}) \boldsymbol{W} \boldsymbol{\epsilon}(\frac{\boldsymbol{W}^T \boldsymbol{q}}{2}) + \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\alpha}) \boldsymbol{W} \boldsymbol{\epsilon}(0) \tag{28}$$

I still need to solve a slight variation of the above problem. Hopefully, the solution can be found in a similar way.

$$f(\boldsymbol{z}) = \frac{a}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + \frac{b}{2}\|\boldsymbol{P}\boldsymbol{z} + (1-\mu)\boldsymbol{P}\boldsymbol{y} - (2-\mu)\boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{29}$$

$$\arg\min_{\boldsymbol{z}} f(\boldsymbol{z}) \tag{30}$$

$$\boldsymbol{z}_{\text{approx}} = \boldsymbol{x} + \frac{b}{a+b}((2-\mu)\boldsymbol{P}\boldsymbol{s} - (1-\mu)\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{x}) \tag{31}$$

To prevent derivations from falling off the page:

$$\boldsymbol{r} = (2-\mu)\boldsymbol{P}\boldsymbol{s} - (1-\mu)\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{x} \tag{32}$$

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|\boldsymbol{P}(\boldsymbol{x}+\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z})+(1-\mu)\boldsymbol{P}\boldsymbol{y}-(2-\mu)\boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{33}$$

$$\boldsymbol{\epsilon}(\Delta\boldsymbol{z}) = \boldsymbol{P}(\boldsymbol{x}+\frac{b}{a+b}((2-\mu)\boldsymbol{P}\boldsymbol{s}-(1-\mu)\boldsymbol{P}\boldsymbol{y}-\boldsymbol{P}\boldsymbol{x})+\Delta\boldsymbol{z})-\boldsymbol{P}\boldsymbol{x}-\frac{b}{a+b}((2-\mu)\boldsymbol{P}\boldsymbol{s}-(1-\mu)\boldsymbol{P}\boldsymbol{y}-\boldsymbol{P}\boldsymbol{x}) \tag{34}$$

$$\boldsymbol{\epsilon}(\Delta\boldsymbol{z}) = \boldsymbol{P}(\boldsymbol{x}+\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}) - \boldsymbol{P}\boldsymbol{x} - \frac{b}{a+b}\boldsymbol{r} \tag{35}$$

$$\boldsymbol{P}(\boldsymbol{x}+\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}) = \boldsymbol{P}\boldsymbol{x} + \frac{b}{a+b}\boldsymbol{r} + \boldsymbol{\epsilon}(\Delta\boldsymbol{z}) \tag{36}$$

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|\boldsymbol{P}\boldsymbol{x}+\frac{b}{a+b}\boldsymbol{r}+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})+(1-\mu)\boldsymbol{P}\boldsymbol{y}-(2-\mu)\boldsymbol{P}\boldsymbol{s}\|_2^2 \tag{37}$$

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|-\boldsymbol{r}+\frac{b}{a+b}\boldsymbol{r}+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})\|_2^2 \tag{38}$$

$$f(\boldsymbol{z}_{\text{approx}}+\Delta\boldsymbol{z}) = \frac{a}{2}\|\frac{b}{a+b}\boldsymbol{r}+\Delta\boldsymbol{z}\|_2^2+\frac{b}{2}\|-\frac{a}{a+b}\boldsymbol{r}+\boldsymbol{\epsilon}(\Delta\boldsymbol{z})\|_2^2 \tag{39}$$

The important distinction here is that μ prevents $\boldsymbol{r}$ from being quantized by $\boldsymbol{P}$'s quantization. So, rounding in the "right direction" could go too far.