Проксимальный оператор

Рассмотрим оптимизационную задачу:

$$g(v) = \mathrm{argmin}_u \frac{1}{2}\|u - v\|^2$$

Решение очевидно, $g(v) = v$.

Усложним задачу, добавив слагаемое:

$$P_f(v) = \mathrm{argmin}_u \frac{1}{2}\|u - v\|^2 + f(u)$$

Это и есть проксимальный оператор.
Идеологически это обобщение проекции.
"Важное свойство":

$$P_f(v) = p \leftrightarrow 0 \in p - v + \partial f(p) \leftrightarrow v - p \in \partial f(p)$$

Примеры и свойства проксимальных операторов

Наиболее распростроённые проксимальные операторы (индикатор условия здесь равен 0, если выполнено, и $+\infty$, если не выполнено).

1. $f(u) = \lambda \|u\|_1 \to P_f(v) = \max(|v| - \lambda, 0)\,\mathrm{sign}\, v$
2. $f(u) = \lambda \|\mathrm{diag}(\sigma_1, \dots, \sigma_n) u\|_1 \to (P_f(v))_i = \max(|v|_i - \sigma_i \lambda_i, 0)\,\mathrm{sign}\, v_i$
3. $f(u) = I\{u \in C\} \to P_f(v) = \mathrm{argmin}_{u \in C} \frac{1}{2}\|u - v\|^2$
4. $f(u) = I\{\|u\|_\infty \leq \lambda\} \to P_f(v) = (P_f(v))_i = \min(1, \frac{\lambda}{|x_i|})x_i$

Применение в задачах оптимизации

Часто в машинном обучении возникает такая оптимизационная задача:

$$\min L(x) + R(x),$$

где $L(x)$ – хорошая, понятная, гладкая, выпуклая функция, а $R(x)$ – регуляризационная добавка, зачастую негладкая.
Из свойств выпуклых функций следует:

$$x = \operatorname{argmin}_x L(x) + R(x) \leftrightarrow 0 \in \partial(L(x) + R(x)) \leftrightarrow$$

$$\leftrightarrow 0 \in \nabla L(x) + \partial R(x) \leftrightarrow -\eta\nabla L(x) \in \eta\partial R(x) \leftrightarrow$$

$$\leftrightarrow (x - \eta\nabla L(x)) - x \in \eta\partial R(x) \leftrightarrow x = P_{\eta R}(x - \eta\nabla L(x))$$

Последняя эквивалентность получена по "Важному свойству". Таким образом, задача свелась к поиску неподвижной точки.

Итерационный процесс

Решая методом простых итераций получаем:

$$x_{n+1} = P_{\eta R}(x_n - \eta \nabla L(x))$$

здесь η имеет смысл шага градиентного спуска, а $P_{\eta R}$ — какой-то фиксированный проксимальный оператор для функции ηR, примеры были на слайде 2.

Используя полученные знания, ответьте на следующие вопросы:

1. В чём схожесть и в чём отличие с обычным градиентным шагом для l_1 регуляризации?
2. Так почему же l_1 регуляризациия приводит к занулению?