Proximal Methods

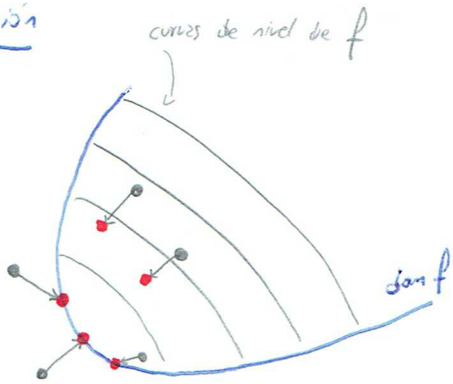
Sea $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ convexa y propia $(f \not\equiv +\infty)$

Definimos el operador proximal $\text{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n$ como

$$\text{prox}_{\lambda f}(z) = \underset{x}{\text{argmin}} \left(f(x) + \frac{1}{2\lambda} \|x - z\|_2^2 \right).$$

La función $f(x) + \frac{1}{2\lambda}\|x - z\|_2^2$ es estrictamente convexa, por lo que el mínimo es único, y la función $\text{prox}_{\lambda f}(z)$ está bien definida.

Interpretación



El parámetro λ regula qué tanto se mueve hacia el mínimo de f, y cuánto se queda cerca de z.

Obs: Si f es la función indicatriz de un conjunto C cerrado y convexo:

$$f(x) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C \end{cases}$$

, entonces el operador proximal coincide con la proyección a C.

Propiedad (Separabilidad de la suma)

Si $f(x,y) = \varphi(x) + \psi(y) \implies \text{prox}_f(v,w) = (\text{prox}_\varphi(v), \text{prox}_\psi(w))$

En particular si $f(x) = \sum_{i=1}^n f_i(x_i) \implies (\text{prox}_f(v))_i = \text{prox}_{f_i}(v_i)$

Prop (del punto fijo)

x^* minimiza f sii $x^* = \text{prox}_f(x^*)$ (consideremos $\lambda = 1$ sin pérdida de generalidad)

Dem

$(\Rightarrow)$ x^* minimiza f $\Rightarrow$ $f(x) \geq f(x^*)$ $\forall x$ $\Rightarrow$ $f(x) + \frac{1}{2}\|x - x^*\|_2^2 \geq f(x^*) + \frac{1}{2}\|x^* - x^*\|_2^2$ $\forall x$

$\Rightarrow$ x^* minimiza $f(x) + \frac{1}{2}\|x - x^*\|_2^2$ $\Rightarrow$ $x^* = prox_f(x^*)$

$(\Leftarrow)$ y minimiza $f(x) + \frac{1}{2}\|x - z\|_2^2$ (es decir $y = prox_f(z)$) sii $0 \in \partial f(y) + (y - z)$

tomando $y = z = x^*$ tenemos que $0 \in \partial f(x^*)$ $\Rightarrow$ x^* minimiza f

Entonces buscar el mínimo de f es equivalente a buscar puntos fijos del operador proximal.

Si $prox_f$ fuera una contracción $\Rightarrow$ $x^{k+1} = prox_f(x^k)$ converge a un punto fijo.

Sin embargo $prox_f$ no es necesariamente una contracción.

Veremos más adelante que de todas maneras tenemos resultados de convergencia.

Ejemplo

Tomemos $f(x) = |x|$ y calculemos $prox_f(v)$

La condición de optimalidad es $0 \in \partial f(x) + x - v$ $\iff$ $v \in x + \partial f(x)$

Dado v, para calcular $x = prox_f(v)$ tenemos que encontrar el x que verifique la cond. de optimalidad $v \in x + \partial f(x)$

Entonces, si $v > 1 \Rightarrow x = v - 1$

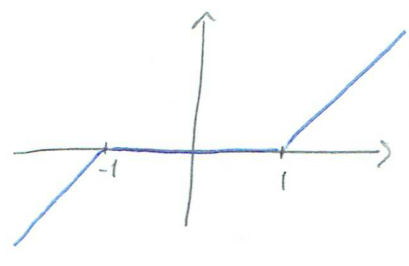
si $v < -1 \Rightarrow x = v + 1$

si $|v| < 1 \Rightarrow x = 0$

o sea $prox_f(v) = \begin{cases} v - 1 & \text{si } v > 1 \\ 0 & \text{si } |v| < 1 \\ v + 1 & \text{si } v < -1 \end{cases}$

Por la prop. de separabilidad, si $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$, entonces

$$(prox_f(v))_i = \begin{cases} v_i - 1 & \text{si } v_i > 1 \\ 0 & \text{si } |v_i| < 1 \\ v_i + 1 & \text{si } v_i < -1 \end{cases}$$



Interpretación: Gradient Flow

Consideremos la ecuación diferencial $\dot{x}(t) = -\nabla f(x(t))$ (flujo de gradiente)

Los puntos de equilibrio son ceros de ∇f, que es donde se minimiza f.

Discretizando la ec. dif. (método Forward-Euler):

$$\frac{x^{k+1} - x^k}{h} = -\nabla f(x^k) \implies x^{k+1} = x^k - h\nabla f(x^k)$$

que es el método de descenso por gradiente con paso h.

Ahora, si usamos Backward-Euler:

$$\frac{x^{k+1} - x^k}{h} = -\nabla f(x^{k+1})$$

que tiene mejores propiedades de convergencia en ec. dif.

resulta $x^{k+1} + h\nabla f(x^{k+1}) = x^k$

que es exactamente la condición de optimalidad para $x^{k+1} = prox_{hf}(x^k)$

Es decir, si discretizamos la ec. dif con Forward-Euler obtenemos el método de descenso por gradiente, y si discretizamos con Backward-Euler obtenemos el método proximal.

Proximal Gradient Method

Consideremos ahora el problema $\min_x f(x) + g(x)$

con $f: \mathbb{R}^n \to \mathbb{R}$ y $g: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ propias y convexas, y f diferenciable.

Usualmente se usa para partir el problema en una parte diferenciable y otra que no necesariamente lo es.

El Proximal Gradient Method es: $x^{k+1} = \text{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$

Es decir, es como hacer un descenso por gradiente de la parte diferenciable, y luego aplicar el operador proximal de g.

Obs:

- Cuando $g \equiv 0$, obtenemos el clásico descenso por gradiente.
- Cuando g es la indicatriz de un conjunto convexo, obtenemos el Projected Gradient Descent.

Interpretación

Como antes, estamos buscando puntos fijos. En este caso $x^* = \text{prox}_{\lambda g}(x^* - \lambda \nabla f(x^*))$. Veamos que este punto fijo es la solución de $\min_x f(x) + g(x)$:

$$x^* = \text{prox}_{\lambda g}(x^* - \lambda \nabla f(x^*)) \iff x^* \text{ minimiza } g(x) + \frac{1}{2\lambda}\| x - x^* + \lambda \nabla f(x^*)\|_2^2$$

$$\iff 0 \in \partial g(x^*) + \frac{1}{\lambda}(x^* - x^* + \lambda \nabla f(x^*)) \iff 0 \in \partial g(x^*) + \nabla f(x^*)$$

$$\iff x^* \text{ minimiza } f(x) + g(x).$$

Obs En este caso, como hay un paso de gradiente, λ no puede ser arbitrario. Por ejemplo si ∇f es Lipschitz de constante L, entonces λ tiene que estar en $[0, \frac{2}{L})$. Se puede hacer line-search también, o Armijo.