3.7 Weight Decay

- **Motivation** :

↳ Alters: (1) Collect more training data → Costly, time consuming
→ Assume already high-quality data, focus the tools.
(2) Monomials, degree d, as d grow bigger, K vars.
monomials of degree d = $\binom{K-1+d}{K-1}$ dramatically increase.

⇒ Often need more fine-grained tool
for adjusting func complexity.

↳ operates by restricting the values that the params can take.
Measure complexity of linear func: $f(x) = w^T x$ by ℓ_p norm

⇒ Objective: Minimize the sum of the prediction loss and the penalty term. ($\|w\|^2$)

$$L(w,b) + \frac{\lambda}{2}\|w\|^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(w^T x^{(i)} + b - y^{(i)}\right)^2 + \left(\frac{\lambda}{2}\|w\|^2\right.$$

↳ ≥ 0 hyperparam fit using validation data

- Minibatch stochastic descent updates for l_2

$$w \leftarrow (1-\eta\lambda)w - \frac{\eta}{|B|}\sum_{i\in B} x^{(i)}\left(w^T x^{(i)} + b - y^{(i)}\right)$$