- $\min_{w} \frac{1}{N} \ell(w^T x_n, y_n) := f(w)$ linear model
- $\min_{w} \frac{1}{N} \ell(f_w(x_n), y_n) := f(w)$ general model
- $w \leftarrow w - \eta \nabla f(w)$

$$f(w) = \frac{1}{N} \sum_{n=1}^{N} f_n(w)$$

$$\nabla f(w) = \frac{1}{N} \sum_{n=1}^{N} \nabla f_n(w)$$

- can be slow when $N \uparrow\uparrow\uparrow \rightarrow$ speed can be slow
- Q: what is the approximate gradient

soln

- Stochastic sampling
  - small subset $B \subseteq \{1, \dots, N\}$
  - estimated gradient

$$\nabla f(w) \approx \frac{1}{B} \sum_{n \in B} \nabla f_n(w)$$

$|B|$: batch size (not fixed re-sample @ each iteration)

- Stochastic Gradient Descent

input: training data $\{x_n, y_n\}_{n=1}^{N}$
initialize w (zero or random)

for $t = 1, 2 \dots$
- sample a small batch $B \subseteq \{1, \dots, N\}$
- update parameter

$$w \leftarrow w - \eta^t \frac{1}{|B|} \sum_{n \in B} \nabla f_n(w)$$

extreme case $|B| = 1$ (each iteration : updated by gradient + zero-mean noise

- stepsize cannot be fixed

- $\eta_t \nabla_w f(w_t) \longrightarrow 0$

noise

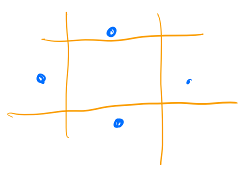
want to approach to zero, hence, make η_t converge to zero.

step decay (different way of decay rate)

- for $f(x)$ that $\nabla f(x)$ is not continuous we can gradient by defining sub-gradient.

- why perceptron learning sometimes does not coverage

check perceptron learning video



problems for data that is not linear-seperable

Momentum

- use previous gradient info

$$V_t = \beta V_t + (1-\beta)\nabla f(W_t)$$

$$W_{t+1} = W_t - \alpha V_t$$

$\beta \in [0, 1)$: discount factors

α : step size

$$\therefore \quad V_t = (1-\beta)\nabla f(W_t) + \beta(1-\beta)\nabla f(W_{t-1}) + \beta^2(1-\beta)\nabla(W_{t-2}) + \cdots$$

Momentum gradient descent

α : learning rate

β = discount factor ($\beta = 0$, no momentum)