Recap:

minimize $f_0(x)$ subject to

$$f_i(x) \leq 0, \quad i = 1, 2, \cdots m$$
$$h_i(x) = 0, \quad i = 1, 2, \cdots p$$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

$$p^* = \min_{x} \max_{\substack{\lambda_i \geq 0 \\ \nu_i}} L(x, \lambda, \nu)$$

$$d^* = \max_{\substack{\lambda_i \geq 0 \\ \nu_i}} \min_{x} L(x, \lambda, \nu)$$

KKT conditions:

If $p^* = d^*$, then if x^* be primal optimal and ~~d^*~~ (λ^*, ν^*) be dual optimal, then

$$\nabla_x f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0$$

$$f_i(x^*) \leq 0$$
$$h_i(x^*) = 0$$
$$\lambda_i^* \geq 0$$
$$\lambda_i^* f_i(x^*) = 0$$

If f_i are convex and h_i are affine, and $\tilde{x}, \tilde{\lambda}, \tilde{v}$ are any points that satisfy KKT conditions

then $\tilde{x}$ ~~is~~ and $(\tilde{\lambda}, \tilde{v})$ are primal and dual optimal with zero duality gap.

SVM:

Let $y_i = f(\bar{x}(i))$

$$\text{minimize } \frac{1}{2} \|w\|_2^2$$

$$\text{subject to } y_i(w \cdot \bar{x}(i) + w_0) \geq 1, \quad i = 1, 2, \dots n$$

Dual:

$$\text{Maximize } \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i \lambda_i (\bar{x}(i) \cdot \bar{x}(j)) y_j \lambda_j$$

$$\text{Subject to } \sum_{i=1}^{n} \lambda_i y_i = 0, \quad \lambda_i \geq 0 \ \ i = 1, 2, \dots n.$$

Moreover

$$w^* = \sum_{i=1}^{n} \lambda_i^* y_i \bar{x}(i)$$

$$w_0^* = y_i - w^* \cdot \bar{x}(i) \text{ for any } i \text{ s.t. } \lambda_i^* > 0$$

Gradient Descent:

we want to minize $f(\underline{x})$

Starting from $\underline{x}_0$

~~Taylor Series Expansion:~~

$f(\underline{x}_0 + h\underline{u})$, $\underline{u}$ unit vector

$h > 0$ small step

Taylor Series expansion:

$\cancel{f(x_0) \approx}$

$$\underbrace{f(\underline{x_0} + h\underline{u})}_{\text{minimize}} \approx f(x_0) + h\underline{u}^T \nabla_x f(x_0)$$

$$u_0 = -\frac{\nabla_x f(x_0)}{\|\nabla_x f(x_0)\|}$$

$$\underline{x_1} = x_0 + h\underline{u_0}$$

Initialize $\underline{x_0}$

Repeat: ~~until~~

$$\underline{x}_{t+1} = x_t - r \nabla f_x(\underline{x}_t)$$

Stop if $\|\nabla_x f(\underline{x}_{t+1})\| \leq \eta$

Stochastic Gradient descent:

Suppose objective function can be written as sum over training examples:

$$MSE(\underline{\omega}) = \sum_{i=1}^{n} (y_i - \omega \cdot \bar{x}(i))^2$$
$$= \sum_{i=1}^{n} g_i(\omega)$$

Initialize $\underline{w_o}$:
Shuffle data $\bar{x}(1), \cdots \bar{x}(n)$
for $i = 1, 2, \cdots n$:
$$\omega := \omega - r \nabla_\omega g_i(\omega)$$

Advantages:
- Faster than gradient descent
- online learning

Disadvantages
- ~~might not converge to local minimum~~
  - performs frequent updates with high variance, hence might keep overshooting.

$\boxed{8.30}$