Support Vector Machine

Our training data consists of N pairs $(X_1, y_1), (X_2, y_2), \dots, (X_N, y_N)$, with $X_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. Define a hyperplane by

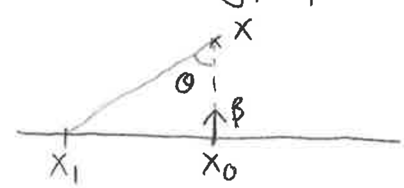
$$\{X : f(X) = X^T\beta + \beta_0 = 0\},$$

with β is a unit vector: $\|\beta\| = 1$. A classification rule is

$$G(X) = \text{sign}\left[X^T\beta + \beta_0\right]$$

Note that $f(X)$ gives the signed distance from a point X to the hyperplane

let X_0 be the nearest point to X on the plane, and X_1 is another point on the plane



$$\therefore \quad X_0^T\beta + \beta_0 = 0 = X_1^T\beta + \beta_0$$

$$\Rightarrow \quad (X_1 - X_0)^T\beta = 0 \qquad (\because \beta \text{ is orthogonal to the plane})$$

Consider $$|(X - X_1)\cdot\beta| = \|X - X_1\|\,\|\beta\|\cos\theta$$
$$= \|X - X_0\|$$

$$\therefore \quad |X^T\beta - X_1^T\beta| = |X^T\beta + \beta_0| = |f(X)| = \|X - X_0\|$$

We can find a function $f(X) = X^T\beta + \beta_0$ with $y_i f(X_i) > 0 \;\; \forall i$ if the points can be separated by a hyperplane. We want to find the hyperplane that creates the biggest margin between the training points for class 1 and -1. The optimization problem is

$$\max_{\beta, \beta_0, \|\beta\|=1} M$$

$$\text{subject to } y_i^T(X_i^T\beta + \beta_0) \geq M, \qquad i = 1, \dots, N$$

Let $\gamma = \frac{\beta}{M}$, $\gamma_0 = \frac{\beta_0}{M}$, then $\|\gamma\| = \frac{\|\beta\|}{M} = \frac{1}{M}$ $(\because \max M \Leftrightarrow \min \|\gamma\|)$

and $$y_i^T(X_i^T\gamma + \gamma_0) \geq 1. \text{ Therefore}$$

The problem can be rephrased as

$$\min_{\beta, \beta_0} \|\beta\|$$

Subject to $y_i(X_i^T\beta + \beta_0) \geq 1, \quad i = 1, \dots, N$

Suppose now that the classes overlap. One way to deal with the overlap is to still maximize M, but allow for some points to be on the wrong side of the margin. Define the slack variables $\xi = (\xi_1, \dots, \xi_N)$

Consider $\min \|\beta\|$

Subject to $y_i^T(X_i^T\beta + \beta_0) \geq 1 - \xi_i \quad \forall i$

$\xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq \text{constant}$

Computationally it is convenient to re-express in the equivalent form

(1) $$\min_{\beta, \beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \qquad (\text{for some } C > 0)$$

Subject to $\xi_i \geq 0, \quad y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i \quad \forall i$

To solve this minimization problem, we consider the Lagrange function

$$L(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(X_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i$$

Define the Lagrange dual function to be

$$g(\alpha, \mu) = \min_{\beta, \beta_0, \xi} L(\beta, \beta_0, \xi, \alpha, \mu)$$

Note that for any β, β_0, ξ satisfies $\xi_i \geq 0$, $y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i$

Given that $\alpha_i \geq 0$ and $\mu_i \geq 0$, we have

$$\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \geq L(\beta, \beta_0, \xi, \alpha, \mu) \geq \min_{\beta, \beta_0, \xi} L(\beta, \beta_0, \xi, \alpha, \mu) = g(\alpha, \mu)$$

Therefore, let $\beta^*, \xi^*, \beta_0^*$ be the solution of (1), the lower bound of $\frac{1}{2}\|\beta^*\|^2 + C\sum_{i=1}^{N}\xi_i^*$ can be found by solving $\max g(\alpha, \mu)$

subject to $\alpha_i \geq 0, \ \mu_i \geq 0 \quad \forall i$

To compute $g(\alpha, \mu) = \min_{\beta, \beta_0, \xi} L(\beta, \beta_0, \xi, \alpha, \mu)$, we minimize L with respect to β, β_0 and ξ. Setting

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \quad \beta - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \quad \Rightarrow \beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \quad C - \alpha_i - \mu_i = 0 \quad \Rightarrow \alpha_i = C - \mu_i \quad \forall i$$

$$\therefore g(\alpha, \mu) = \frac{1}{2} \left(\sum_{i=1}^{N} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{N} \alpha_j y_j x_j\right) + \sum_{i=1}^{N} \alpha_i \xi_i - \left(\sum_{i=1}^{N} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{N} \alpha_j y_j x_j\right) + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

And thus we get a dual problem

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{subject to } 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

Let $\hat{\alpha}$ be the maximizer. It is known that $\frac{1}{2}\|\beta^*\|^2 + C \sum_{i=1}^{N} \xi_i^* = g(\hat{\alpha})$

($\because$ Slater's constraint qualification : $\because$ the original problem is convex and strictly feasible)

Therefore the solution $\beta^* = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$

Note that by KKT conditions

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0 \qquad \mu_i \xi_i = 0$$

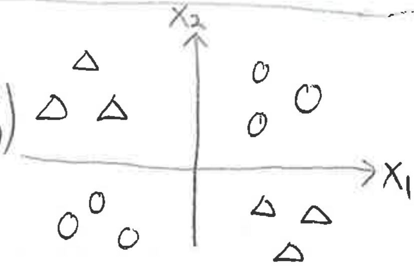
$\Rightarrow$ if $\hat{\alpha}_i > 0$, then $y_i(x_i^T\beta + \beta_0) = 1 - \xi_i$. The corresponding x_i's are called the support vectors, as $\hat{\beta}$ is represented in term of them.

Among these support points, some will lie on the edge of the margin ($\xi_i^* = 0$). Any of these margin points ($\hat{\alpha}_i > 0$, $\hat{\xi}_i = 0$) can be used to solve for β_0

We typically use an average of all the solutions for numerical stability. Maximizing the dual problem can be done by quadratic programming. Given $\hat{\alpha}$, and thus $\hat{\beta}$ and $\hat{\beta}_0$, the decision function can be written as

$$\hat{G}(x) = \operatorname{sign}[\hat{f}(x)] = \operatorname{sign}[x^T\hat{\beta} + \hat{\beta}_0] = \operatorname{sign}\left[\sum_{i=1}^N \hat{\alpha}_i y_i x_i^T x + \hat{\beta}_0\right]$$

How to find a hyperplane to separate the data $\{\vec{X}_i = (X_{i1}, X_{i2})\}$ with $y_i = \begin{cases} 1 & \text{if } \operatorname{sign}(X_{i1}) \neq \operatorname{sign}(X_{i2}) \quad (\text{e.g. } (+,-), (-,+)) \\ -1 & \text{if } \operatorname{sign}(X_{i1}) = \operatorname{sign}(X_{i2}) \quad (\text{e.g. } (+,+), (-,-)) \end{cases}$

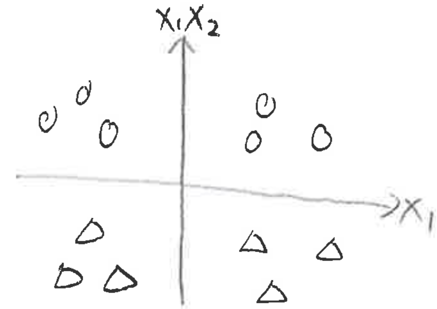


No good choice. However, if we consider the transformation

$$h(\vec{X}_i) = h(X_{i1}, X_{i2}) = (X_{i1}, X_{i1}X_{i2})$$

then

$$\begin{aligned} (+,-) &\to (+,-) \\ (+,+) &\to (+,+) \\ (-,-) &\to (-,+) \\ (-,+) &\to (-,-) \end{aligned}$$



Now the data can be separated by $X_1X_2 = 0$

In general, consider the hyperplane $f(x) = h(x)^T\beta + \beta_0$

As before, β and β_0 can be estimated by first solving the dual problem

$$\max_{\alpha} \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j \langle h(x_i), h(x_j)\rangle$$

$$\text{subject to} \quad 0 \le \alpha_i \le C, \quad \sum_{i=1}^N \alpha_i y_i = 0$$

After getting $\hat{\alpha}$, we can compute $\hat{\beta}$ and $\hat{\beta}_0$, and then the decision function

$$\hat{G}(x) = \operatorname{sign}\left[\sum_{i=1}^N \hat{\alpha}_i y_i \langle h(x_i), h(x)\rangle + \hat{\beta}_0\right]$$

Notice that we need not specify the transformation $h(x)$ at all, but require only knowledge of the kernel function

$$K(x, x') = \langle h(x), h(x')\rangle$$

which is a dot-product function.

Mercer's condition A real-valued function $K(x,y)$ is said to fulfill Mercer's condition if for all square-integrable functions $g(x)$ (i.e. $\int g^2(x)dx < \infty$), $\iint g(x)K(x,y)g(y)\,dx\,dy \geq 0$

For a given $K(x,y)$, there exists transformation $h(x)$ such that

$$K(x,y) = \langle h(x), h(y) \rangle$$

if and only if $K(x,y)$ satisfies Mercer's condition

Note that if $K_1(x,y)$ and $K_2(x,y)$ satisfy Mercer's condition, then for any $\theta_1 \geq 0$ and $\theta_2 \geq 0$,

$$\iint g(x)[\theta_1 K_1(x,y) + \theta_2 K_2(x,y)]g(y)\,dx\,dy \qquad \forall g(x), \int g^2(x)dx < \infty$$

$$= \theta_1 \iint g(x)K_1(x,y)g(y)\,dx\,dy + \theta_2 \iint g(x)K_2(x,y)g(y)\,dx\,dy \geq 0$$

$\therefore$ $\theta_1 K_1 + \theta_2 K_2$ is also a kernel function

---

Example : For $K(x,y) = c > 0$

$$\iint g(x)K(x,y)g(y)\,dx\,dy = c\iint g(x)g(y)\,dx\,dy$$

$$= c\left(\int g(x)dx\right)\left(\int g(y)dy\right)$$

$$= c\left(\int g(x)dx\right)^2 \geq 0$$

$\therefore$ $K(x,y)$ satifies Mercer's condition ($h(x) = \sqrt{c}$)

Example : Consider $\vec{x} = (x_1, x_2)$ $\quad K(\vec{x},\vec{y}) = (\vec{x}^T\vec{y})^2 = (x_1y_1 + x_2y_2)^2$

Note that $K(\vec{x},\vec{y}) = (x_1y_1)^2 + 2(x_1y_1)(x_2y_2) + (x_2y_2)^2$

$$\therefore \iint g(x)K(x,y)g(y)\,dx\,dy = \left(\int g(x)x_1^2dx\right)\left(\int g(y)y_1^2dy\right)$$

$$+ 2\left(\int g(x)x_1x_2dx\right)\left(\int g(y)y_1y_2dy\right)$$

$$+ \left(\int g(x)x_2^2dx\right)\left(\int g(y)y_2^2dy\right) \geq 0$$

Example: In general, for $\vec{x} \in \mathbb{R}^d$, $K(\vec{x},\vec{y}) = (\vec{x}^T\vec{y})^p = (\sum_{i=1}^{d} x_i y_i)^p$

$$K(\vec{x},\vec{y}) = \sum_{r_1+r_2+\dots+r_d=p} (x_1y_1)^{r_1}(x_2y_2)^{r_2}\cdots(x_dy_d)^{r_d}\frac{p!}{r_1!\,r_2!\cdots r_d!}$$

$$\therefore \iint g(x)K(x,y)g(y)\,dx\,dy = \sum_{r_1+\dots+r_d=p}\frac{p!}{r_1!\cdots r_d!}\left(\int g(x)x_1^{r_1}\cdots x_d^{r_d}dx\right)\left(\int g(y)y_1^{r_1}\cdots y_d^{r_d}dy\right)$$

$$\geq 0$$

Three popular choices for K in the SVM literature

dth-Degree polynomial: $K(x,x') = (1+\langle x,x'\rangle)^d \quad d\geq 0$

Radial basis: $K(x,x') = e^{(-\gamma\|x-x'\|^2)}$

Neural network (Sigmoid function): $K(x,x') = \tanh(K_1\langle x,x'\rangle + K_2)$

where $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

Example: Consider $\vec{x} = (x_1, x_2)$ $\quad K(\vec{x},\vec{y}) = (1+\langle\vec{x},\vec{y}\rangle)^2$

$$\begin{aligned} K(\vec{x},\vec{y}) &= (1+x_1y_1+x_2y_2)^2 \\ &= 1+2x_1y_1+2x_2y_2+x_1^2y_1^2+x_2^2y_2^2+2x_1y_1x_2y_2 \\ &= (1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,x_2^2,\sqrt{2}x_1x_2)^T(1,\sqrt{2}y_1,\sqrt{2}y_2,y_1^2,y_2^2,\sqrt{2}y_1y_2) \end{aligned}$$

$\therefore$ We have $h(\vec{x}) = h(x_1,x_2) = (1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,x_2^2,\sqrt{2}x_1x_2)^T$

The SVM as a Penalization Method

With $f(x) = h(x)^T\beta + \beta_0$, consider the optimization problem

$$\min_{\beta_0,\beta}\sum_{i=1}^{N}[1-y_if(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2$$

with $[x]_+ = \max(x,0)$. Let β_0^*, β^* be the solution of this optimization problem. It can be showed that $(\beta_0^*, \beta^*, \xi^*)$ is also the solution of

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \ , \text{ s.t. } \xi_i \geq 0, \quad y_i f(x_i) \geq 1-\xi_i \ \forall_i$$

with $\lambda = \frac{1}{C}$.

Pf: Note that $\xi_i \geq 0$ and $\xi_i \geq 1 - y_i f(x_i)$

$$\Leftrightarrow \ \xi_i \geq [1-y_i f(x_i)]_+ = \max(0, 1-y_i(h(x_i)^T\beta+\beta_0))$$

Clearly, if $(\tilde{\beta}, \tilde{\beta_0}, \tilde{\xi})$ is the solution of

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \ , \quad \xi_i \geq [1-y_i(h(x_i)^T\beta+\beta_0)]_+$$

then $\tilde{\xi}_i = [1-y_i(h(x_i)^T\tilde{\beta}+\tilde{\beta}_0)]_+$

$\therefore$ The minimization problem is equivalent to

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}[1-y_i f(x_i)]_+$$

It is known that if $f_1, \dots, f_m$ are convex functions, then $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ is also convex.

Since $1 - y_i h(x_i)^T\beta - y_i\beta_0$ and 0 are convex (linear) in β and β_0, $[1-y_i f(x_i)]_+ = \max(1-y_i(h(x_i)^T\beta+\beta_0), 0)$ is also convex

Therefore $\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}[1-y_i f(x_i)]_+$ is a convex optimization problem.

By considering $h(x)^T\beta + \beta_0 = (h(x) \ \ 1)^T\begin{pmatrix}\beta\\ \beta_0\end{pmatrix} = \tilde{h}(x)^T\tilde{\beta}$

where $\tilde{h}(x) : x \to \begin{pmatrix}h(x)\\ 1\end{pmatrix}$ and $\tilde{\beta} = \begin{pmatrix}\beta\\ \beta_0\end{pmatrix}$, we can consider $\beta_0 = 0$

Then the minimization can be further simplified as $\min_{\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} L_i(\beta)$

where $L_i(\beta) = [1-y_i h(x_i)^T\beta]_+$

Let $g(\beta) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} L_i(\beta)$, we minimize $g(\beta)$ by gradient descent

Starting from β_0, update

$$\beta_{t+1} \leftarrow \beta_t - \eta \nabla g(\beta_t)$$

Consider $\nabla g(\beta) = \beta + C \sum_{i=1}^{N} \nabla_\beta L_i(\beta)$ $\qquad \nabla_\beta L_i(\beta) = \begin{cases} -y_i h(x_i) & \text{if } 1 > y_i h(x_i)^T \beta \\ 0 & \text{if } y_i h(x_i)^T \beta \geq 1 \end{cases}$

$\therefore$ We have $\beta_{t+1} = \beta_t - \eta(\beta_t + C \sum_{i=1}^{N} \nabla L_i(\beta_t))$

Computing each update takes $O(N)$ time. N is the size of training dataset

---

Stochastic Gradient Descent (SGD)

Instead of evaluating gradient over all examples, we evaluate it for one training example each time

$$\beta_{t+1} = \beta_t - \eta(\beta_t + C \nabla L_i(\beta_t))$$
$$= (1-\eta)\beta_t - \eta C \nabla L_i(\beta_t)$$

index "i" is randomly chosen from the training set

How to choose η?

The advantage of SGD is quick update in each iteration. It is reasonable to adjust η (e.g. getting smaller when β_t is close to the solution β_*), but the update of η should not be complicated. In particular, choosing η such that $G(\eta) = g((1-\eta)\beta_t - \eta C \nabla L_i(\beta_t))$ is minimum is not desirable.

We first start from usual gradient descent update $\beta_{t+1} = \beta_t - \eta \nabla g(\beta_t)$

For the solution β_*, we have $\nabla g(\beta_*) = 0$. Note that $\nabla^2 g(\beta) = I \quad \forall \beta \setminus (y_i h(x_i)^T \beta = 1)$

Here, to be more general, we assume $\mu I \leq \nabla^2 f(x) \leq L I$

i.e. $\mu \|v\|^2 \leq v^T \nabla^2 f(x) v \leq L \|v\|^2$ for some $L > \mu > 0$

then
$$\beta_{t+1} - \beta_* = \beta_t - \beta_* - \eta(\nabla g(\beta_t) - \nabla g(\beta_*))$$
$$= \beta_t - \beta_* - \eta \nabla^2 g(\xi_t)(\beta_t - \beta_*)$$
$$= (I - \eta \nabla^2 g(\xi_t))(\beta_t - \beta_*)$$

for some ξ_t between β_t and β_*

(Suppose ∇g is cts between β_t, β_*)

$\therefore \|\beta_{t+1} - \beta_*\| \leq \|I - \eta \nabla^2 g(\xi_t)\| \, \|\beta_t - \beta_*\| \leq \max(|1-\eta\mu|, |1-\eta L|) \, \|\beta_t - \beta_*\|$

9

Now, consider $\beta_{t+1} = \beta_t - \eta \nabla g(\beta_t)$, with $\nabla g(\beta_t) = \sum_{i=1}^{N} \nabla g_i(\beta_t)$

For $\nabla g(\beta) = \beta + C \sum_{i=1}^{N} \nabla_\beta L_i(\beta) = \sum_{i=1}^{N} \left(\frac{\beta}{N} + C \nabla_\beta L_i(\beta) \right)$, $\nabla g_i(\beta) = \frac{\beta}{N} + C \nabla_\beta L_i(\beta)$

Assume each observation $(\vec{x}_i, y_i)$ are independent, then if we randomly choose $i_t \in \{i=1,..,N\}$, then $E(\nabla g_{i_t}(\beta_t)) = \frac{1}{N} \nabla g(\beta_t)$

Now consider $\beta_{t+1} = \beta_t - \eta \nabla g_{i_t}(\beta_t)$

$$\beta_{t+1} - \beta_* = \beta_t - \beta_* - \eta \left(\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \right) - \frac{\eta}{N} \left(\nabla g(\beta_t) - \nabla g(\beta_*) \right)$$

$$= \left(I - \frac{\eta}{N} \nabla^2 g(\xi_t) \right)(\beta_t - \beta_*) - \eta \left(\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \right)$$

We handle $\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t)$ by variance $Var(\beta_{t+1,j} - \beta_{*,j} | \beta_t)$ where $\beta_{t+1,j}$ and $\beta_{*,j}$ is the jth entry of β_{t+1} and β_* respectively

$$Var(\beta_{t+1,j} - \beta_{*,j} | \beta_t)$$

$$= Var\left[\eta \left(\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \right)_j | \beta_t \right] \quad \left(\because \left(I - \frac{\eta}{N} \nabla^2 g(\xi_t) \right)(\beta_t - \beta_*) \text{ is fixed if } \beta_t \text{ is given} \right)$$

$$= \eta^2 E\left[\left(\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \right)_j^2 | \beta_t \right]$$

$$\therefore E\left[(\beta_{t+1} - \beta_*)_j^2 | \beta_t \right] = \left[E((\beta_{t+1} - \beta_*)_j | \beta_t) \right]^2 + Var\left[(\beta_{t+1} - \beta_*)_j | \beta_t \right]$$

$$= \left[E(\beta_{t+1} - \beta_* | \beta_t)_j \right]^2 + \eta^2 E\left[\left(\nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \right)_j^2 | \beta_t \right]$$

Summation over all entries j, we get

$$E\left[\|\beta_{t+1} - \beta_*\|^2 | \beta_t \right] = \| E[\beta_{t+1} - \beta_* | \beta_t] \|^2 + \eta^2 E\left[\| \nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \|^2 | \beta_t \right]$$

$$\le (1 - \eta\mu)^2 \|\beta_{t+1} - \beta_*\|^2 + \eta^2 M$$

Here we assume $\eta \ll 1$ and $E\left[\| \nabla g_{i_t}(\beta_t) - \frac{1}{N} \nabla g(\beta_t) \|^2 | \beta_t \right] \le M \quad \forall \beta_t, i_t$

By tower property, we have

$$E\left[\|\beta_{t+1} - \beta_*\|^2 \right] \le (1 - \eta\mu)^2 E\left[\|\beta_t - \beta_*\|^2 \right] + \eta^2 M \quad (\text{assume } 0 < \eta\mu < 1)$$

$$\le (1 - \eta\mu)^2 \left((1 - \eta\mu)^2 E\left[\|\beta_{t-1} - \beta_*\|^2 \right] + \eta^2 M \right) + \eta^2 M$$

$$\vdots$$

$$\le \eta^2 M \left(1 + (1 - \eta\mu)^2 + (1 - \eta\mu)^4 + \dots \right)$$

$$= \frac{\eta^2 M}{1 - (1 - \eta\mu)^2} = \frac{\eta M}{2\mu - \eta\mu^2}$$

This phenomenon is called converging to a noise ball. Rather than approaching the optimum, SGD (with a constant step size) converges to a region around the optimum. This is okay for applications that only need approximate solutions.

The bound $\frac{\eta M}{2\mu - \eta\mu^2} \approx 0$ if we choose a very small η. However, small η at the beginning implies slow convergence. We want to decrease η_t when t increases.

Consider
$$E[\|\beta_{t+1} - \beta_*\|^2] \le (1-\eta_t\mu)^2 E[\|\beta_t - \beta_*\|^2] + \eta_t^2 M$$
$$\le (1-\eta_t\mu) E[\|\beta_t - \beta_*\|^2] + \eta_t^2 M \quad (\text{for } \eta_t\mu < 1)$$

Minimize right-hand side by setting $\nabla_\eta = 0$

$$-\mu E[\|\beta_t - \beta_*\|^2] + 2\eta_t M = 0 \Rightarrow \eta_t = \frac{\mu}{2M} E[\|\beta_t - \beta_*\|^2]$$

let $\rho_t = E[\|\beta_t - \beta_*\|^2]$, for $\eta_t = \frac{\mu}{2M}\rho_t$

$$\rho_{t+1} \le \left(1 - \frac{\mu}{2M}\rho_t\mu\right)\rho_t + \left(\frac{\mu}{2M}\rho_t\right)^2 M$$
$$= \rho_t - \frac{\mu^2}{2M}\rho_t^2 + \frac{\mu^2}{4M}\rho_t^2 = \rho_t - \frac{\mu^2}{4M}\rho_t^2 = \rho_t\left(1 - \frac{\mu^2}{4M}\rho_t\right)$$

$$\Rightarrow \frac{1}{\rho_{t+1}} \ge \frac{1}{\rho_t}\left(1 - \frac{\mu^2}{4M}\rho_t\right)^{-1} \quad \left(\text{assume } \frac{\mu^2}{4M}\rho_t < 1\right.$$
$$\ge \frac{1}{\rho_t}\left(1 + \frac{\mu^2}{4M}\rho_t\right). \quad \text{then for } 1 - z^2 < 1 \Rightarrow 1 + z < \frac{1}{1-z}$$
$$\text{if } 1 - z > 0 \left.\right)$$
$$= \frac{1}{\rho_t} + \frac{\mu^2}{4M} \ge \cdots \ge \frac{\mu^2(t+1)}{4M} + \frac{1}{\rho_0}$$

$\Rightarrow \rho_t \to 0$ as $t \to \infty$

Since $\rho_t \le \frac{1}{\frac{1}{\rho_0} + \frac{\mu^2 t}{4M}} = \frac{4M\rho_0}{4M + \mu^2\rho_0 t}$, we choose $\eta_t = \frac{\mu}{2M}\left(\frac{4M\rho_0}{4M + \mu^2\rho_0 t}\right)$

$$= \frac{2\mu\rho_0}{4M + \mu^2\rho_0 t}$$

General form is $\alpha_t = \frac{\alpha_0}{1 + \gamma t}$