MIT 9.520/6.7910
Statistical Learning Theory and Applications

Class 04: Logistic Regression and SGD

Lorenzo Rosasco

Beyond least squares regression

$$(y - f(x))^2 \mapsto \ell(y, f(x)).$$

or rather

$$\ell(yf(x))$$

for classification

In particular:

- Logistic loss $\log(1 + e^{-yf(x)})$
- Hinge loss $|1 - yf(x)|_+ = \max\{1 - yf(x), 0\}$.

Other loss can be used, e.g. exponential loss (boosting).

Outline

Max margin, minimal norm and regularization

Logistic regression, GD and SGD

Step-size choice and convergence for GD

Step-size choice and convergence for SGD

Support vector machines and subgradients

ERM for classification

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w) + \lambda \|w\|^2, \qquad \lambda \geq 0.$$

- Logistic loss $\mapsto$ (regularized) logistic regression.
- Hinge $\mapsto$ SVM.

Regularization and minimal norm separating solution

Let

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w) + \lambda \|w\|^2, \qquad \lambda \geq 0.$$

and

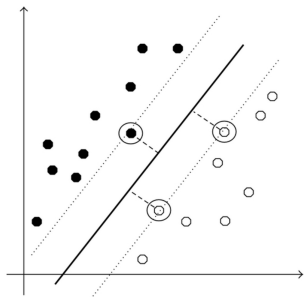
$$\hat{v}^\dagger = \arg\max_{w \in \mathbb{R}^d} \min_{i=1,\ldots,n} y_i x_i^\top w \qquad \text{s.t.} \qquad \|w\| = 1.$$

For both the logistic and hinge loss

$$\lim_{\lambda \to 0} \frac{\widehat{w}_\lambda}{\|\widehat{w}_\lambda\|} = \hat{v}^\dagger.$$

Max margin and minimal norm

$$\hat{v}^\dagger = \arg\max_{w \in \mathbb{R}^d} \min_{i=1,\dots,n} y_i x_i^\top w \qquad \text{s.t.} \qquad \|w\| = 1.$$



Minimal norm and max margin

$$\hat{v}^\dagger = \arg\max_{w \in \mathbb{R}^d} \min_{i=1,\dots,n} y_i x_i^\top w \qquad \text{s.t.} \qquad \|w\| = 1.$$

Let

$$\widehat{w}^\dagger = \arg\min_{w \in \mathbb{R}^d} \|w\| \qquad \text{s.t.} \qquad y_i x_i^\top w \geq 1, \qquad i = 1, \dots, n.$$

Minimal norm and max margin

$$\hat{v}^\dagger = \arg\max_{w \in \mathbb{R}^d} \min_{i=1,\dots,n} y_i x_i^\top w \qquad \text{s.t.} \qquad \|w\| = 1.$$

Let

$$\widehat{w}^\dagger = \arg\min_{w \in \mathbb{R}^d} \|w\| \qquad \text{s.t.} \qquad y_i x_i^\top w \geq 1, \qquad i = 1, \dots, n.$$

Then

$$\frac{\widehat{w}^\dagger}{\|\widehat{w}^\dagger\|} = \hat{v}^\dagger$$

and

$$\widehat{w}^\dagger = \hat{v}^\dagger \cdot \min_{i=1,\dots,n} y_i x_i^\top \hat{v}^\dagger.$$

Proof of equivalence

By definition $\widehat{w}^\dagger$ solves

$$\min \|w\| \quad \text{s.t.} \quad y_i x_i^\top w \geq 1,$$

so that $\hat{v}^\dagger = \frac{\widehat{w}^\dagger}{\left\|\widehat{w}^\dagger\right\|}$ satisfies

$$\min_i y_i x_i^\top \hat{v}^\dagger \quad = \quad \frac{1}{\left\|\widehat{w}^\dagger\right\|}.$$

Proof of equivalence

By definition $\widehat{w}^\dagger$ solves

$$\min \|w\| \quad \text{s.t.} \quad y_i x_i^\top w \geq 1,$$

so that $\hat{v}^\dagger = \frac{\widehat{w}^\dagger}{\|\widehat{w}^\dagger\|}$ satisfies

$$\min_i y_i x_i^\top \hat{v}^\dagger = \frac{1}{\|\widehat{w}^\dagger\|}.$$

Thus

$$\max_{\|v\|=1} \min_i y_i x_i^\top v = \frac{1}{\|\widehat{w}^\dagger\|}.$$

Proof of equivalence

By definition $\widehat{w}^\dagger$ solves

$$\min \|w\| \quad \text{s.t.} \quad y_i x_i^\top w \geq 1,$$

so that $\hat{v}^\dagger = \frac{\widehat{w}^\dagger}{\|\widehat{w}^\dagger\|}$ satisfies

$$\min_i y_i x_i^\top \hat{v}^\dagger = \frac{1}{\|\widehat{w}^\dagger\|}.$$

Thus

$$\max_{\|v\|=1} \min_i y_i x_i^\top v = \frac{1}{\|\widehat{w}^\dagger\|}.$$

Hence

$$\frac{\widehat{w}^\dagger}{\|\widehat{w}^\dagger\|} = \hat{v}^\dagger, \qquad \widehat{w}^\dagger = \hat{v}^\dagger \cdot \min_i y_i x_i^\top \hat{v}^\dagger.$$

Remarks

- Interpolation implies separation:

$$w^\top x_i = y_i \quad \Leftrightarrow \quad y_i w^\top x_i = 1 \quad \Rightarrow \quad y_i w^\top x_i \geq 1.$$

Remarks

- Interpolation implies separation:

$$w^\top x_i = y_i \quad \Leftrightarrow \quad y_i w^\top x_i = 1 \quad \Rightarrow \quad y_i w^\top x_i \geq 1.$$

- Hence overparametrization ($n < d$) implies separation.

Remarks

- Interpolation implies separation:

$$w^\top x_i = y_i \quad \Leftrightarrow \quad y_i w^\top x_i = 1 \quad \Rightarrow \quad y_i w^\top x_i \geq 1.$$

- Hence overparametrization ($n < d$) implies separation.

- The proof extends to homogeneous function $f(ax) = af(x)$ for $a > 0$ (e.g. ReLU nets).

Max margin and regularization

$$\hat{v}^\dagger = \arg\max_{w \in \mathbb{R}^d} \min_{i=1,\dots,n} y_i x_i^\top w \qquad \text{s.t.} \qquad \|w\| = 1.$$

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w) + \lambda \|w\|^2, \qquad \lambda \geq 0.$$

Regularization allows to handle the trade-off between separating the data and having simple solutions.

Outline

From regularization to optimization

Problem For $\lambda >$, solve

$$\min_{w \in \mathbb{R}^d} \widehat{L}(w) + \lambda \|w\|^2,$$

where

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w).$$

Minimization

Assume ℓ convex and continuous, let

$$\widehat{L}_\lambda(w) = \widehat{L}(w) \; + \; \lambda \|w\|^2.$$

- Coercive[1], continuous, strongly convex functional
  $\Rightarrow$ a minimizer exists and is unique.

- Computations depends on the considered loss.

[1] $\lim_{\|w\| \to \infty} \widehat{L}_\lambda(w) = \infty$.

Logistic regression

$$\widehat{L}_\lambda(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top w}) + \lambda \|w\|^2.$$

- $\widehat{L}_\lambda$ is smooth

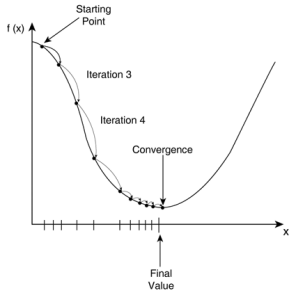
$$\nabla \widehat{L}_\lambda(w) = -\frac{1}{n}\sum_{i=1}^{n} \frac{x_i y_i}{1 + e^{y_i x_i^\top w}} + 2\lambda w.$$

- Optimality conditions give nonlinear equations

$$\nabla \widehat{L}_\lambda(w) = 0,$$

  so we use gradient methods.

Gradient descent



Let $F : \mathbb{R}^d \to \mathbb{R}$ differentiable,

$$w_0 = 0, \quad w_{t+1} = w_t - \eta \nabla F(w_t),$$

Gradient descent for logistic regression

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top w}) + \lambda \|w\|^2.$$

Consider

$$w_{t+1} = w_t - \frac{1}{B} \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{x_i y_i}{1 + e^{y_i w_t^\top x_i}} + 2\lambda w_t \right),$$

with $B = \sup_{i=1,\dots,n} \|x_i\| + 2\lambda$.

Complexity

Time: $O(ndt)$ for n examples, d dimension, t steps.

SGD for logistic regression

Gradient descent

$$w_{t+1} = w_t - \frac{1}{B}\left(-\frac{1}{n}\sum_{i=1}^{n}\frac{x_i y_i}{1+e^{y_i w_t^\top x_i}} + 2\lambda w_t\right)$$

Stochastic gradient (descent)

$$w_{t+1} = w_t - \gamma\left(-\frac{x_{i_t} y_{i_t}}{1+e^{y_{i_t} w_t^\top x_{i_t}}} + 2\lambda w_t\right)$$

Complexity

$O(ndt) \rightsquigarrow O(dt)$

SGD flavors

$$w_{t+1} = w_t - \gamma \left(-\frac{x_{i_t} y_{i_t}}{1 + e^{y_{i_t} w_t^\top x_{i_t}}} + 2\lambda w_t \right)$$

Common variations

- Mini-batching: use $b \in (1, n)$ points to compute a gradient estimate.
- Acceleration.

What about step size choice and covergence

Gradient descent

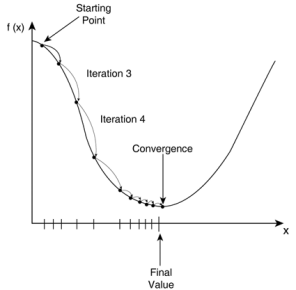
$$w_{t+1} = w_t - \frac{1}{B}\left(-\frac{1}{n}\sum_{i=1}^{n}\frac{x_i y_i}{1+e^{y_i w_t^\top x_i}} + 2\lambda w_t\right)$$

Stochastic gradient (descent)

$$w_{t+1} = w_t - \gamma\left(-\frac{x_{i_t} y_{i_t}}{1+e^{y_{i_t} w_t^\top x_{i_t}}} + 2\lambda w_t\right)$$

Outline

Gradient descent



Let $F : \mathbb{R}^d \to \mathbb{R}$ differentiable, convex and such that

$$\|\nabla F(w) - \nabla F(w')\| \leq B\|w - w'\|$$

(e.g. $\sup_w \|\underbrace{\nabla^2 F(w)}_{\text{Hessian}}\| \leq B$)

Then

$$w_0 = 0, \quad w_{t+1} = w_t - \frac{1}{B}\nabla F(w_t),$$

converges to a minimizer w^* of F.

Function value rates

Rates can be obtained under different assumptions. We consider two case:

Convex + B-smooth:

$$F(w_t) - F(w^\star) \leq \frac{B\|w_0 - w^\star\|^2}{2t}.$$

Polyak–Łojasiewicz (PL) + B-smooth:

$$F(w_t) - F(w^\star) \leq \left(1 - \tfrac{\mu}{B}\right)^t \big(F(w_0) - F(w^\star)\big).$$

GD rate: convex + smooth

If F is convex and B-smooth, then gradient descent with step size $\eta = \frac{1}{B}$ satisfies

$$F(w_t) - F(w^\star) \leq \frac{B\|w_0 - w^\star\|^2}{2t}.$$

Proof steps:

1. Descent lemma
2. Convexity
3. Combining
4. Telescoping
5. Final bound

Descent lemma

If F is B-smooth, then for all u, v, the following quadratic bound holds

$$F(v) \leq F(u) + \nabla F(u)^\top (v - u) + \frac{B}{2}\|v - u\|^2.$$

Descent lemma

If F is B-smooth, then for all u, v, the following quadratic bound holds

$$F(v) \leq F(u) + \nabla F(u)^\top (v - u) + \frac{B}{2}\|v - u\|^2.$$

(Hint) Apply the fundamental theorem of calculus

$$h(1) = h(0) + \int_0^1 h'(a)da$$

to

$$h(a) = F(u + a(v - u)).$$

Descent lemma (cont.)

$$F(v) \;\leq\; F(u) + \nabla F(u)^\top (v-u) + \tfrac{B}{2}\|v-u\|^2.$$

In particular, for GD $w_{t+1} = w_t - \frac{1}{B}\nabla F(w_t)$, taking $v = w_{t+1}$, $u = w_t$

$$F(w_{t+1}) \;\leq\; F(w_t) - \tfrac{1}{2B}\|\nabla F(w_t)\|^2.$$

Convexity

If F is convex, then for all u, v,

$$F(u) - F(v) \;\leq\; \nabla F(u)^\top (u - v).$$

In particular, with $u = w_t$ and $v = w^\star$:

$$F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining

Descent lemma and convexity

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2, \qquad F(w_t) \leq F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star).$$

Combining

Descent lemma and convexity

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2, \qquad F(w_t) \leq F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$F(w_{t+1}) \leq F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

Combining

Descent lemma and convexity

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2, \qquad F(w_t) \leq F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$F(w_{t+1}) \leq F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

Identity:

$$\nabla F(w_t)^\top (w_t - w^\star) - \frac{1}{2B}\|\nabla F(w_t)\|^2 = \frac{B}{2}\left(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\right).$$

Combining

Descent lemma and convexity

$$F(w_{t+1}) \;\leq\; F(w_t) - \tfrac{1}{2B}\|\nabla F(w_t)\|^2, \qquad F(w_t) \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$F(w_{t+1}) \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \tfrac{1}{2B}\|\nabla F(w_t)\|^2.$$

Identity:

$$\nabla F(w_t)^\top (w_t - w^\star) - \tfrac{1}{2B}\|\nabla F(w_t)\|^2 = \tfrac{B}{2}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Then

$$F(w_{t+1}) \;\leq\; F(w^\star) + \tfrac{B}{2}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Telescoping

$$F(w_{t+1}) \;\leq\; F(w^\star) + \tfrac{B}{2}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Summing for $t = 0, \ldots, T-1$:

$$\sum_{t=0}^{T-1} \big(F(w_{t+1}) - F(w^\star)\big) \;\leq\; \tfrac{B}{2}\|w_0 - w^\star\|^2.$$

Final bound

Since $F(w_t)$ is nonincreasing,

$$T\left(F(w_T) - F(w^\star)\right) \leq \sum_{t=0}^{T-1} \left(F(w_t) - F(w^\star)\right).$$

Therefore

$$F(w_T) - F(w^\star) \leq \frac{B}{2T}\|w_0 - w^\star\|^2.$$

□

Function value rates

Rates can be obtained under different assumptions. We consider two case:

Convex + B-smooth:

$$F(w_T) - F(w^\star) \leq \frac{B\|w_0 - w^\star\|^2}{2T}.$$

Polyak–Łojasiewicz (PL) + B-smooth:

$$F(w_t) - F(w^\star) \leq \left(1 - \tfrac{\mu}{B}\right)^t \big(F(w_0) - F(w^\star)\big).$$

Polyak–Łojasiewicz (PL) condition

A differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the **PL** condition with parameter μ if $\forall w \in \mathbb{R}^d$,

$$\frac{1}{2}\|\nabla F(w)\|^2 \;\geq\; \mu\left(F(w) - F(w^\star)\right).$$

Strong convexity implies PL

$$F(w^\star) \;\geq\; F(w) + \nabla F(w)^\top (w^\star - w) + \tfrac{\mu}{2}\|w - w^\star\|^2 \quad \Rightarrow \quad F(w) - F(w^\star) \;\leq\; \tfrac{1}{2\mu}\|\nabla F(w)\|^2.$$

Indeed, rearranging

$$F(w) - F(w^\star) \;\leq\; \nabla F(w)^\top (w - w^\star) - \tfrac{\mu}{2}\|w - w^\star\|^2.$$

Strong convexity implies PL

$$F(w^\star) \;\geq\; F(w) + \nabla F(w)^\top (w^\star - w) + \tfrac{\mu}{2}\|w - w^\star\|^2 \quad \Rightarrow \quad F(w) - F(w^\star) \;\leq\; \tfrac{1}{2\mu}\|\nabla F(w)\|^2.$$

Indeed, rearranging

$$F(w) - F(w^\star) \;\leq\; \nabla F(w)^\top (w - w^\star) - \tfrac{\mu}{2}\|w - w^\star\|^2.$$

By Cauchy–Schwarz,

$$\nabla F(w)^\top (w - w^\star) \;\leq\; \|\nabla F(w)\|\,\|w - w^\star\|,$$

Strong convexity implies PL

$$F(w^\star) \;\geq\; F(w) + \nabla F(w)^\top (w^\star - w) + \tfrac{\mu}{2}\|w - w^\star\|^2 \quad \Rightarrow \quad F(w) - F(w^\star) \;\leq\; \tfrac{1}{2\mu}\|\nabla F(w)\|^2.$$

Indeed, rearranging

$$F(w) - F(w^\star) \;\leq\; \nabla F(w)^\top (w - w^\star) - \tfrac{\mu}{2}\|w - w^\star\|^2.$$

By Cauchy–Schwarz,

$$\nabla F(w)^\top (w - w^\star) \;\leq\; \|\nabla F(w)\|\,\|w - w^\star\|,$$

and by Young's inequality ($ab \leq \frac{1}{2\varepsilon}a^2 + \frac{\varepsilon}{2}b^2$, $\varepsilon = \mu$)

$$\|\nabla F(w)\|\,\|w - w^\star\| \;\leq\; \tfrac{1}{2\mu}\|\nabla F(w)\|^2 + \tfrac{\mu}{2}\|w - w^\star\|^2.$$

GD rate: PL + smooth

If F is B-smooth and satisfies PL, then gradient descent with step size $\eta = \frac{1}{B}$ satisfies

$$F(w_t) - F(w^\star) \leq \left(1 - \frac{\mu}{B}\right)^t (F(w_0) - F(w^\star)).$$

Proof steps:

1. Descent lemma
2. PL
3. Combine
4. Iterate

PL proof is short!

From descent lemma:

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

PL proof is short!

From descent lemma:

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \geq 2\mu\,(F(w_t) - F(w^\star)).$$

PL proof is short!

From descent lemma:

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \geq 2\mu\,(F(w_t) - F(w^\star)).$$

Therefore

$$F(w_{t+1}) - F(w^\star) \leq \left(1 - \frac{\mu}{B}\right)(F(w_t) - F(w^\star)).$$

PL proof is short!

From descent lemma:

$$F(w_{t+1}) \leq F(w_t) - \frac{1}{2B}\|\nabla F(w_t)\|^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \geq 2\mu\,(F(w_t) - F(w^\star)).$$

Therefore

$$F(w_{t+1}) - F(w^\star) \leq \left(1 - \frac{\mu}{B}\right)(F(w_t) - F(w^\star)).$$

Iterating:

$$F(w_t) - F(w^\star) \leq \left(1 - \frac{\mu}{B}\right)^t (F(w_0) - F(w^\star)).$$

□

Remarks

- Constant step-size is nice.
- Adaptive choices can be considered (line search, variable metric).
- But also other approximations (quasi-Newton or SGD, next).

Outline

SGD: finite-sum case

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w).$$

GD:

$$w_{t+1} = w_t - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla f_i(w_t).$$

SGD: finite-sum case

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w).$$

GD:

$$w_{t+1} = w_t - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla f_i(w_t).$$

SGD (sample $i_t \sim \{1, \dots, n\}$):

$$w_{t+1} = w_t - \eta_t \nabla f_{i_t}(w_t).$$

SGD: expectation case

$$F(w) = \mathbb{E}_{\xi}[f(w;\xi)].$$

Draw $\xi_t \sim \mathcal{D}$:

$$w_{t+1} = w_t - \eta_t \nabla f(w_t;\xi_t).$$

SGD: expectation case

$$F(w) = \mathbb{E}_{\xi}[f(w;\xi)].$$

Draw $\xi_t \sim \mathcal{D}$:

$$w_{t+1} = w_t - \eta_t \nabla f(w_t;\xi_t).$$

Unbiasedness:

$$\mathbb{E}[\nabla f(w_t;\xi_t) \mid w_t] = \nabla F(w_t).$$

Finite-sum as special case

Let $\xi \sim \text{Unif}\{1, \ldots, n\}$ and $f(w; \xi = i) = f_i(w)$.

Then

$$F(w) = \mathbb{E}_\xi[f(w; \xi)] = \frac{1}{n} \sum_{i=1}^{n} f_i(w).$$

The finite-sum case is exactly the expectation case with discrete ξ!

Variance assumption

We assume bounded variance

$$\mathbb{E}\left[\|\nabla f(w;\xi) - \nabla F(w)\|^2 \mid w\right] \leq \sigma^2.$$

Together with unbiasedness, we can write

$$g_t = \nabla F(w_t) + \zeta_t,$$

with $\mathbb{E}[\zeta_t \mid w_t] = 0, \quad \mathbb{E}[\|\zeta_t\|^2 \mid w_t] \leq \sigma^2$

Variance assumption

We assume bounded variance

$$\mathbb{E}\left[\|\nabla f(w;\xi) - \nabla F(w)\|^2 \mid w\right] \;\leq\; \sigma^2.$$

Together with unbiasedness, we can write

$$g_t = \nabla F(w_t) + \zeta_t,$$

with $\mathbb{E}[\zeta_t \mid w_t] = 0, \quad \mathbb{E}[\|\zeta_t\|^2 \mid w_t] \leq \sigma^2$ (equivalently, $\mathbb{E}[g_t \mid w_t] = \nabla F(w_t)$ and $\mathrm{Var}(g_t \mid w_t) = \mathbb{E}[\|g_t - \nabla F(w_t)\|^2] \leq \sigma^2$).

Function value rates: SGD

Rates can be obtained under different assumptions. We consider two cases:

Convex + B-smooth ("diminishing" steps and averaging):

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{B\|w_0 - w^\star\|^2 + \sigma^2}{2\sqrt{T}}.$$

Polyak–Łojasiewicz (PL) + B-smooth (constant step):

$$\mathbb{E}[F(w_t)] - F(w^\star) \;\leq\; \left(1 - \tfrac{\mu}{B}\right)^t (F(w_0) - F(w^\star)) + \frac{B\eta\sigma^2}{2\mu}.$$

SGD rate: convex + smooth

If F is convex, B-smooth, with bounded variance stochastic gradients, then SGD average

$$\bar{w}_T = \frac{1}{T} \sum_{t=0}^{T-1} w_t$$

with steps $\eta_t = \frac{1}{B\sqrt{T}}$ satisfies

$$\mathbb{E}[F(\bar{w}_T) - F(w^*)] \leq \frac{\|w_0 - w^*\|^2 + \sigma^2}{2\sqrt{T}}.$$

Proof steps:

1. One-step inequality (smoothness + SGD update)
2. Combining with convexity
3. Telescoping
4. Averaging

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

By B-smoothness:

$$F(w_{t+1}) \leq F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{B\eta_t^2}{2}\|g_t\|^2.$$

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

By B-smoothness:

$$F(w_{t+1}) \leq F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{B\eta_t^2}{2}\|g_t\|^2.$$

Taking conditional expectation:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \eta_t\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\,\mathbb{E}[\|g_t\|^2 \mid w_t].$$

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

By B-smoothness:

$$F(w_{t+1}) \leq F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{B\eta_t^2}{2}\|g_t\|^2.$$

Taking conditional expectation:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \eta_t\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\,\mathbb{E}[\|g_t\|^2 \mid w_t].$$

Variance decomposition:

$$\mathbb{E}[\|g_t\|^2 \mid w_t] = \|\nabla F(w_t)\|^2 + \mathrm{Var}(g_t \mid w_t) \leq \|\nabla F(w_t)\|^2 + \sigma^2.$$

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

By B-smoothness:

$$F(w_{t+1}) \leq F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{B\eta_t^2}{2}\|g_t\|^2.$$

Taking conditional expectation:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \eta_t\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\,\mathbb{E}[\|g_t\|^2 \mid w_t].$$

Variance decomposition:

$$\mathbb{E}[\|g_t\|^2 \mid w_t] = \|\nabla F(w_t)\|^2 + \mathrm{Var}(g_t \mid w_t) \leq \|\nabla F(w_t)\|^2 + \sigma^2.$$

Derive the inequality

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \left(\eta_t - \frac{B\eta_t^2}{2}\right)\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

One-step inequality for SGD

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

By B-smoothness:

$$F(w_{t+1}) \leq F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{B\eta_t^2}{2}\|g_t\|^2.$$

Taking conditional expectation:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \eta_t\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\,\mathbb{E}[\|g_t\|^2 \mid w_t].$$

Variance decomposition:

$$\mathbb{E}[\|g_t\|^2 \mid w_t] = \|\nabla F(w_t)\|^2 + \mathrm{Var}(g_t \mid w_t) \leq \|\nabla F(w_t)\|^2 + \sigma^2.$$

Derive the inequality

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \leq F(w_t) - \left(\eta_t - \frac{B\eta_t^2}{2}\right)\|\nabla F(w_t)\|^2 + \frac{B\eta_t^2}{2}\sigma^2.$$

If $\eta_t \leq 1/B$, then $\eta_t - \frac{B\eta_t^2}{2} \geq \frac{\eta_t}{2}$.

Combining

One-step inequality and convexity

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2, \quad F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining

One-step inequality and convexity

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2, \quad F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2.$$

Combining

One-step inequality and convexity

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2, \quad F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2.$$

Identity (with the stochastic gradient g_t):

$$g_t^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|g_t\|^2 = \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Combining

One-step inequality and convexity

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2, \quad F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2.$$

Identity (with the stochastic gradient g_t):

$$g_t^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|g_t\|^2 = \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Conditional expectation and variance decomposition give

$$\nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 = \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \mathbb{E}[\|w_{t+1} - w^\star\|^2 \mid w_t]\big) + \tfrac{\eta_t}{2}\operatorname{Var}(g_t \mid w_t).$$

Combining

One-step inequality and convexity

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2, \quad F(w_t) - F(w^\star) \;\leq\; \nabla F(w_t)^\top (w_t - w^\star).$$

Combining:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w^\star) + \nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta_t^2}{2}\sigma^2.$$

Identity (with the stochastic gradient g_t):

$$g_t^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|g_t\|^2 = \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2\big).$$

Conditional expectation and variance decomposition give

$$\nabla F(w_t)^\top (w_t - w^\star) - \tfrac{\eta_t}{2}\|\nabla F(w_t)\|^2 = \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \mathbb{E}[\|w_{t+1} - w^\star\|^2 \mid w_t]\big) + \tfrac{\eta_t}{2}\operatorname{Var}(g_t \mid w_t).$$

Therefore

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w^\star) + \tfrac{1}{2\eta_t}\big(\|w_t - w^\star\|^2 - \mathbb{E}[\|w_{t+1} - w^\star\|^2 \mid w_t]\big) + \tfrac{\eta(1+B\eta_t)}{2}\sigma^2.$$

Telescoping

For constant step size $\eta_t = \eta$,

$$\mathbb{E}[F(w_{t+1}) - F(w^\star)] \;\leq\; \frac{\mathbb{E}\,\|w_t - w^\star\|^2 - \mathbb{E}\,\|w_{t+1} - w^\star\|^2}{2\eta} + \frac{\eta}{2} \underbrace{(1 + B\eta)}_{\leq 2 \text{ if } \eta \leq 1/B} \sigma^2.$$

Telescoping

For constant step size $\eta_t = \eta$,

$$\mathbb{E}[F(w_{t+1}) - F(w^\star)] \;\leq\; \frac{\mathbb{E}\,\|w_t - w^\star\|^2 - \mathbb{E}\,\|w_{t+1} - w^\star\|^2}{2\eta} + \frac{\eta}{2}\underbrace{(1 + B\eta)}_{\leq 2 \text{ if } \eta \leq 1/B}\sigma^2.$$

Summing for $t = 0, \ldots, T-1$:

$$\sum_{t=1}^{T} \mathbb{E}[F(w_t) - F(w^\star)] \;\leq\; \frac{\|w_0 - w^\star\|^2}{2\eta} + T\eta\sigma^2.$$

Final bound

By convexity, for the averaged iterate $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$:

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[F(w_t) - F(w^\star)].$$

Final bound

By convexity, for the averaged iterate $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$:

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[F(w_t) - F(w^\star)].$$

Therefore

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{\|w_0 - w^\star\|^2}{2\eta T} + \eta\sigma^2.$$

Final bound

By convexity, for the averaged iterate $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$:

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[F(w_t) - F(w^\star)].$$

Therefore

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{\|w_0 - w^\star\|^2}{2\eta T} + \eta\sigma^2.$$

Choosing $\eta = \frac{1}{B\sqrt{T}}$

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{B\|w_0 - w^\star\|^2}{2\sqrt{T}} + \frac{\sigma^2}{B\sqrt{T}}.$$

□

Function value rates: SGD

Rates can be obtained under different assumptions. We consider two cases:

Convex + B-smooth ("diminishing" steps):

$$\mathbb{E}[F(\bar{w}_T) - F(w^\star)] \;\leq\; \frac{\|w_0 - w^\star\|^2 + B\sigma^2}{2\sqrt{T}}.$$

Polyak–Łojasiewicz (PL) + B-smooth (constant step):

$$\mathbb{E}[F(w_t)] - F(w^\star) \;\leq\; \left(1 - \tfrac{\mu}{B}\right)^t (F(w_0) - F(w^\star)) + \frac{B\eta\sigma^2}{2\mu}.$$

SGD rate: PL + smooth

If F is B-smooth and satisfies PL, then SGD with constant step size $\eta \leq \frac{1}{B}$ satisfies

$$\mathbb{E}[F(w_t) - F(w^\star)] \;\leq\; \left(1 - \tfrac{\mu}{2B}\right)^t (F(w_0) - F(w^\star)) + \frac{B\eta\sigma^2}{2\mu}.$$

Proof steps:

1. One-step inequality
2. PL inequality
3. Combine
4. Iterate (geometric decay + variance floor)

PL proof with SGD

From the one-step inequality:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta^2}{2}\sigma^2.$$

PL proof with SGD

From the one-step inequality:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta^2}{2}\sigma^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \;\geq\; 2\mu\,(F(w_t) - F(w^\star)).$$

PL proof with SGD

From the one-step inequality:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta^2}{2}\sigma^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \;\geq\; 2\mu\,(F(w_t) - F(w^\star)).$$

Therefore

$$\mathbb{E}[F(w_{t+1}) - F(w^\star) \mid w_t] \;\leq\; \Big(1 - \eta\mu\Big)(F(w_t) - F(w^\star)) + \tfrac{B\eta^2}{2}\sigma^2.$$

PL proof with SGD

From the one-step inequality:

$$\mathbb{E}[F(w_{t+1}) \mid w_t] \;\leq\; F(w_t) - \tfrac{\eta}{2}\|\nabla F(w_t)\|^2 + \tfrac{B\eta^2}{2}\sigma^2.$$

Apply PL:

$$\|\nabla F(w_t)\|^2 \;\geq\; 2\mu\,(F(w_t) - F(w^\star)).$$

Therefore

$$\mathbb{E}[F(w_{t+1}) - F(w^\star) \mid w_t] \;\leq\; \Big(1 - \eta\mu\Big)(F(w_t) - F(w^\star)) + \tfrac{B\eta^2}{2}\sigma^2.$$

Taking expectations and iterating:

$$\mathbb{E}[F(w_t) - F(w^\star)] \;\leq\; \Big(1 - \eta\mu\Big)^t (F(w_0) - F(w^\star)) + \frac{B\eta\sigma^2}{2\mu}.$$

□

Remarks

- $\eta = O(1/\sqrt{T})$ gives a $\eta = O(1/\sqrt{T})$ bound.
- η_t diminishing at each iteration can be considered (always ensuring convergence!).
- More adaptive choices can be considered.

Outline

Max margin, minimal norm and regularization

Logistic regression, GD and SGD

Step-size choice and convergence for GD

Step-size choice and convergence for SGD

Support vector machines and subgradients

Hinge loss and support vector machines

$$\widehat{L}_\lambda(w) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} |1 - y_i x_i^\top w|_+ \; + \; \lambda \|w\|^2}_{\text{non-smooth \& strongly-convex}}$$

Consider the "left" derivative of $\ell(a) = |1-a|_+$

$$w_{t+1} = w_t - \frac{1}{B\sqrt{t}} \left(\frac{1}{n}\sum_{i=1}^{n} S_i(w_t) \; + \; 2\lambda w_t \right)$$

$$S_i(w) = y_i x_i \ell^-(y_i x_i^\top w) = \begin{cases} -y_i x_i & \text{if } y_i x_i^\top w \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

with $B = \sup_{i=1,\ldots,n} \|x_i\| + 2\lambda$.

Subgradient

Let $F : \mathbb{R}^p \to \mathbb{R}$ convex,

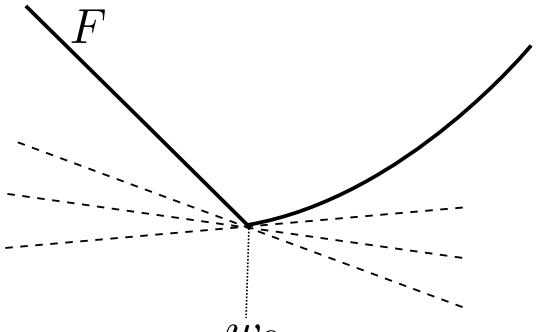
Subgradient

$\partial F(w_0)$ set of vectors $v \in \mathbb{R}^p$ such that, for every $w \in \mathbb{R}^p$

$$F(w) - F(w_0) \geq (w - w_0)^\top v$$

Properties

- In one dimension $\partial F(w_0) = [F'_-(w_0), F'_+(w_0)]$.
- There are analogous to chain rule and sum rule for derivative.



Subgradient method

Let $F : \mathbb{R}^p \to \mathbb{R}$ convex, with subdifferential bounded by B, and $\gamma_t = \frac{1}{B\sqrt{t}}$ then,

$$w_{t+1} = w_t - \gamma_t v_t$$

with $v_t \in \partial F(w_t)$ converges to the minimizer of F.

Note: it is not a descent method.

Subgradient method for SVM

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} |1 - y_i x_i^\top w|_+ \; + \; \lambda \|w\|^2$$

Consider

$$w_{t+1} = w_t - \frac{1}{B\sqrt{t}} \left(\frac{1}{n} \sum_{i=1}^{n} S_i(w_t) \; + \; 2\lambda w_t \right)$$

$$S_i(w_t) = \begin{cases} -y_i x_i & \text{if } y_i x_i^\top w \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Complexity

Time: $O(ndt)$ for n examples, d dimensions, t steps.

SGD for SVM

Subgradient method

$$w_{t+1} = w_t - \frac{1}{B\sqrt{t}} \left(\frac{1}{n} \sum_{i=1}^{n} S_i(w_t) \; + \; 2\lambda w_t \right)$$

Stochastic subgradient method

$$w_{t+1} = w_t - \gamma \left(S_{i_t}(w_t) \; + \; 2\lambda w_t \right)$$

Complexity

$O(ndt) \rightsquigarrow O(dt)$

Connection to the perceptron

- Replace the hinge loss with
$$\ell(yf(x)) = |-yf(x)|_+.$$
- Set $\lambda = 0$.

Reasoning as above we can solve ERM by

$$w_{t+1} = w_t - \frac{1}{B\sqrt{t}} \left(\frac{1}{n} \sum_{i=1}^{n} S_i(w_t) \; + \; 2\lambda w_t \right), \; S_i(w_t) = \begin{cases} -y_i x_i & \text{if } y_i x_i^\top w \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a "batch" version of the perceptron algorithm,

$$w_{t+1} = w_t - \gamma \left(S_t(w_t) \right), \quad S_i(w_t) = \begin{cases} -y_t x_t & \text{if } y_t x^\top w_t \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Representer theorem of SVM

By induction

$$c_{t+1} = c_t - \frac{1}{B\sqrt{t}} \left(\frac{1}{n} \sum_{i=1}^{n} S_i(c_t) + 2\lambda c_t \right)$$

with e_i the i-th element of the canonical basis,

$$f_t(x) = \sum_{i=1}^{n} x^\top x_i (c_t)_i$$

and

$$S_i(c_t) = \begin{cases} -y_i e_i & \text{if } y_i f_t(x_i) < 1 \\ 0 & \text{otherwise} \end{cases}.$$

What else

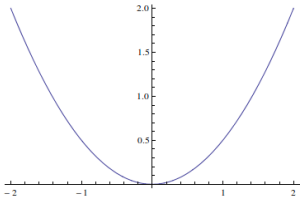
- Why are they called support vector machines?

- And what about the margin and all that?

Optimality condition for SVM

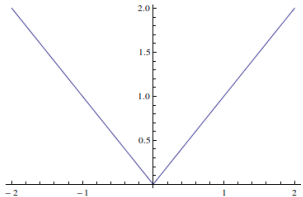
Smooth Convex

$$\nabla F(w_*) = 0$$



Non-smooth Convex

$$0 \in \partial F(w)$$



$$0 \in \partial F(w_*) \quad \Leftrightarrow \quad 0 \in \partial \frac{1}{n} \sum_{i=1}^{n} |1 - y_i x_i^\top w|_+ + \lambda 2w$$

$$\Leftrightarrow \quad w \in -\partial \sum_{i=1}^{n} \frac{1}{n2\lambda} |1 - y_i x_i^\top w|_+.$$

Optimality condition for SVM (cont.)

The optimality condition can be rewritten as

$$0 = \frac{1}{n}\sum_{i=1}^{n}(y_i x_i c_i) + 2\lambda w \quad \Rightarrow \quad w = -\sum_{i=1}^{n} x_i(\frac{y_i c_i}{2\lambda n}).$$

where

$$c_i = c_i(w) \in [\ell^-(y_i w^\top x_i), \ell^+(y_i w^\top x_i)]$$

with ℓ^-, ℓ^+ left, right derivatives of $\ell(a) = |1-a|_+$.

A direct computation gives

$$\begin{aligned} c_i &= -1 \quad \text{if} \quad y_i f(x_i) < 1 \\ -1 \leq c_i &\leq 0 \quad \text{if} \quad y_i f(x_i) = 1 \\ c_i &= 0 \quad \text{if} \quad y_i f(x_i) > 1 \end{aligned}$$

A more familiar form

$$f(x) = -\sum_{i=1}^{n} x^\top x_i \left(\frac{y_i c_i}{2n\lambda}\right)$$

$$\begin{aligned} c_i &= -1 \quad \text{if} \quad yf(x_i) < 1 \\ -1 \leq c_i &\leq 0 \quad \text{if} \quad yf(x_i) = 1 \\ c_i &= 0 \quad \text{if} \quad yf(x_i) > 1 \end{aligned}$$

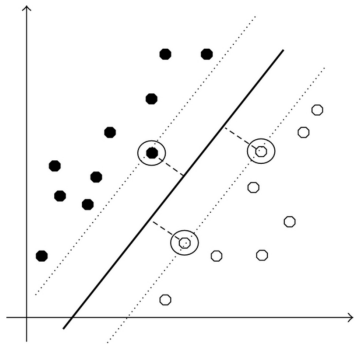
Let $C = \frac{1}{2n\lambda}$ and $\alpha_i = -c_i C$.

$$f(x) = \sum_{i=1}^{n} x^\top x_i y_i \alpha_i$$

$$\begin{aligned} \alpha_i &= C \quad \text{if} \quad yf(x_i) < 1 \\ 0 \leq \alpha_i &\leq C \quad \text{if} \quad yf(x_i) = 1 \\ \alpha_i &= 0 \quad \text{if} \quad yf(x_i) > 1 \end{aligned}$$

Support vectors

$$
\begin{aligned}
c_i = -1 \quad & \text{if} \quad yf(x_i) < 1 \\
-1 \leq c_i \leq 0 \quad & \text{if} \quad yf(x_i) = 1 \\
c_i = 0 \quad & \text{if} \quad yf(x_i) > 1
\end{aligned}
$$



Sparsity and SVM solvers

The conditions

$$\begin{aligned} c_i &= 1 \quad \text{if} \quad y_i f(x_i) < 1 \\ 0 \leq c_i &\leq 1 \quad \text{if} \quad y_i f(x_i) = 1 \\ c_i &= 0 \quad \text{if} \quad y_i f(x_i) > 1 \end{aligned}$$

show that the SVM solution is sparse wrt the training points.

- Classical Quadratic Programming solvers for SVM exploit sparsity.
- Subgradient methods require only matrix vector multiplications, hence are preferable for large scale problems.

Back to the margin

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} |1 - y_i x_i^\top w|_+ \; + \; \lambda \|w\|^2.$$

For $C = \frac{1}{2n\lambda}$, consider the following equivalent formulation

$$\min_{w \in \mathbb{R}^d} C \sum_{i=1}^{n} \xi_i \; + \; \frac{1}{2} \|w\|^2,$$

subj. to for all $i = 1, \ldots, n$,

$$\xi_i \geq 0, \qquad y_i x_i^\top w \geq 1 - \xi_i$$

The *slack* variables ξ_i's quantify how much constraints are violated.

Soft and hard margin SVM

This is the classical *soft margin* SVM formulation

$$\min_{w \in \mathbb{R}^d} C \sum_{i=1}^{n} \xi_i \; + \; \frac{1}{2}\|w\|^2, \quad \text{subj. to} \quad \xi_i \geq 0, \quad y_i x_i^\top w \geq 1 - \xi_i, \; \forall \; i = 1, \ldots, n.$$

The name comes from considering the limit case $C \to \infty$

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2, \qquad \text{subj. to} \quad y_i x_i^\top w \geq 1, \quad \forall \; i = 1, \ldots, n,$$

called *hard margin* SVM (the max margin problem)

Max margin

$$\min_{w \in \mathbb{R}^d} \|w\|^2, \quad \text{subj. to} \quad y_i x_i^\top w \geq 1, \ \forall \ i = 1, \ldots, n.$$

The above problem has a geometric interpretation.

For linearly separable data

- $2/\|w\|$ is the margin: smallest distance of each class to $x^\top w$.
- The constraint select functions linearly separating the data.

Hard margin SVM: find the max margin solution separating the data.

Summing up

- Logistic regression and SVM are instances of penalized ERM.
- Optimization by gradient descent/subgradient method.
- Optimality conditions and support vectors.
- Margin.