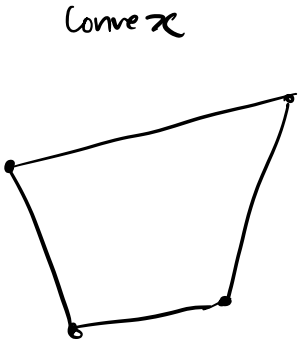
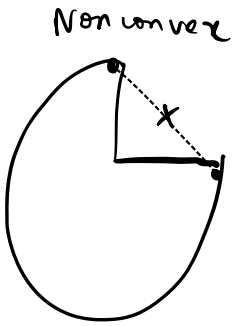
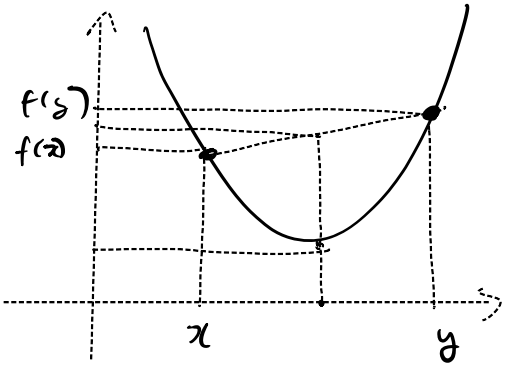
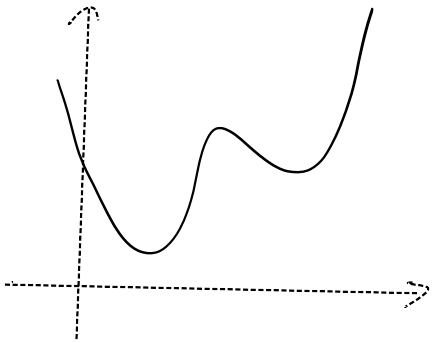
Lecture 15

RECAP **Convex optimization**









Lemma If f is differentiable, then f is convex if & only if domain (f) is convex &

$$f(y) - f(x) \geq \langle y - x, \nabla f(x) \rangle$$

$$f(y) - f(x) \leq \langle y - x, \nabla f(y) \rangle$$

Convex optimization problems can be solved efficiently :)

$$\min_{x \in A} f(x)$$

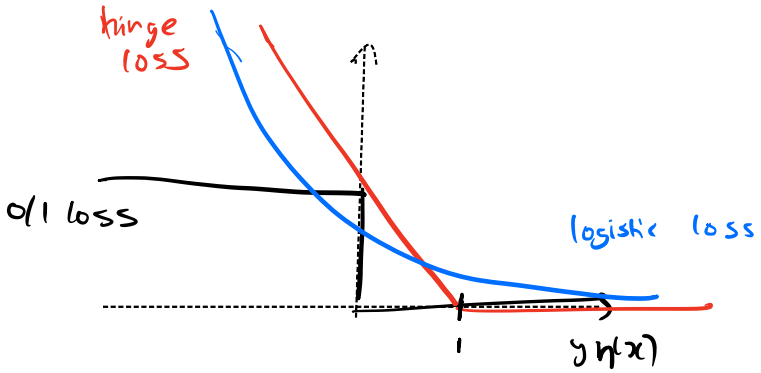
① $f(x)$ is convex ② A is convex.

Let $\mathcal{H} \subseteq \mathbb{R}^d$ parameterized by $w \in \mathcal{H}$

$$ERM_{\mathcal{H}}(S) = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$$

Lemma If ℓ is a convex loss (in terms of w), & $\mathcal{H}$ is convex, then $ERM_{\mathcal{H}}(S)$ is a convex optimization problem.

A common technique to handle a non-convex loss is to instead consider a **convex surrogate**.



hinge loss : $\ell(h,(x,y)) = \max(1 - yh(x), 0)$

logistic loss : $\ell(h,(x,y)) = \log\left(1 + e^{-yh(x)}\right)$

TODAY

* More on convex surrogates
* Algorithms for convex optimization

Convex surrogates

Let $R(h) = \mathbb{E}_{(x,y)\sim D} \mathbb{1}\{yh(x) \leq 0\}$ be 0/1 risk.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be some other function & define

$$R_\phi(h) = \mathbb{E}_{(x,y)\sim D}\left[\phi(yh(x))\right]$$

to be the surrogate risk.

Recall $R^* = \inf_{f: X \to Y} R(f)$ is Bayes-optimal risk.

Let $R^*_\phi = \inf_{f: X \to Y} R_\phi(f)$ denote Bayes-optimal ϕ-risk.

Def We say surrogate loss ϕ is classification-calibrated if for any sequence of functions f_i & every distribution D over (x,y),

$$R_\phi(f_i) \to R^*_\phi \implies R(f_i) \to R^*.$$

Note We want $R_\phi(f_i) \to R_\phi^*$ (This is Bayes-optimal predictor for ϕ-risk). If we optimize over some limited hypothesis class which doesn't the Bayes-optimal predictor for ϕ-risk.

Thm (Bartlett, Jordan, McAuliffe '06) Consider a surrogate loss $\phi(yh(x))$. If ϕ is convex, then it is classification calibrated if & only if $\phi'(0)$ exists & $\phi'(0) < 0$.

Corollary: Hinge loss & logistic loss are classification calibrated.

Note Squared loss $\ell(h, (x,y)) = (h(x) - y)^2$ doesn't fit form of $\phi(yh(x))$, but we can show it is also classification-calibrated.

Gradient descent (GD)

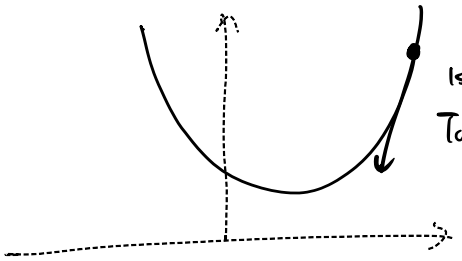
Consider an unconstrained convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

Let $f(x)$ is differentiable

GD:

1. Initialize w_1
2. For $t = 1, \ldots, T$
3. $\quad w_{t+1} \leftarrow w_t - \eta \nabla f(w_t)$



1st-order optimization
Take a step with some step size.

Thm (convergence rate of GD) Let f be a convex, ρ-Lipschitz function i.e.

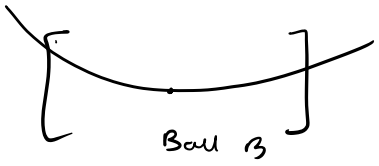
$$|f(x) - f(y)| \leq \rho \|x - y\|_2 \quad \forall \; x, y \in \mathbb{R}^d.$$

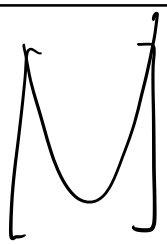
Let $w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} f(w)$. Let $\|w^* - w_1\| \leq B$, where w_1 is initialization for GD. Suppose we run GD for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ & let $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$.

Then $\bar{w}$ satisfies

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

If $T \geq \frac{B^2 \rho^2}{\varepsilon^2}$, $f(\bar{w}) - f(w^*) \leq \varepsilon$.





Proof

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T} \sum_{t=1}^{T} w_t\right) - f(w^*)$$

$$\leq \frac{1}{T} \sum_{t=1}^{T} \left(f(w_t) - f(w^*)\right)$$

$$= \frac{1}{T} \sum_{t=1}^{T} \left(f(w_t) - f(w^*)\right) \quad - (1)$$

Because f is convex, we have

$$f(w_t) - f(w^*) \leq \langle w_t - w^*, \nabla f(w_t) \rangle \quad - (2)$$

Combining (1) & (2),

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, \nabla f(w_t) \rangle \quad - (3)$$

We will upper bound $\sum_{t=1}^{T} \langle w_t - w^*, \nabla f(w_t) \rangle$

Lemma (Iterative update) Let $v_1, \ldots, v_T$ be an arbitary sequence of vectors. Consider any algorithm with an update rule

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

Then

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \leq \frac{\|w^* - w_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

If $\|w^* - w_1\| \leq B$ & $\|v_t\| \leq \rho$ and we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$,

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \leq \frac{B\rho}{\sqrt{T}} .$$

Proof

$$\langle w_t - w^*, v_t \rangle = \frac{1}{\eta} \langle w_t - w^*, \eta v_t \rangle$$

$$= \frac{1}{2\eta} \left(- \|w_t - w^* - \eta v_t\|_2^2 + \|w_t - w^*\|_2^2 + \eta^2 \|v_t\|^2 \right)$$

$$= \frac{1}{2\eta} \left(- \|w_{t+1} - w^*\|^2 + \|w_t - w^*\|^2 \right) + \frac{\eta}{2} \|v_t\|^2$$

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(- \|w_{t+1} - w^*\|^2 + \|w_t - w^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

$$\sum_{t=1}^{T} \left(- \|w_{t+1} - w^*\|^2 + \|w_t - w^*\|^2 \right) = \|w_1 - w^*\|^2 - \|w_{T+1} - w^*\|^2$$

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle = \frac{1}{2\eta} \left(\|w_1 - w^*\|^2 - \|w_{T+1} - w^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

$$\leq \frac{1}{2\eta} \|w_1 - w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

$$\leq \frac{B^2}{2\eta} + \frac{\eta}{2} T \rho^2$$

$$\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \leq \frac{B^2}{2\eta T} + \frac{\eta}{2} \rho^2$$

$$\eta = \sqrt{\frac{B^2}{\rho^2 T}}$$

$$\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \leq \frac{B\rho}{\sqrt{T}}$$

Claim If f is convex & ρ-Lipschitz, then $\|\nabla f(x)\| \leq \rho \ \forall x$.

Proof

$$f(x') - f(x) \geq \langle x' - x, \nabla f(x) \rangle$$

Take $x' = x + \varepsilon \dfrac{\nabla f(x)}{\|\nabla f(x)\|}$

$$f(x') - f(x) \geq \varepsilon \|\nabla f(x)\|$$

By ρ-Lipschitzness $\quad |f(x') - f(x)| \leq \rho \|x - x'\|$

$$\leq \rho \varepsilon$$

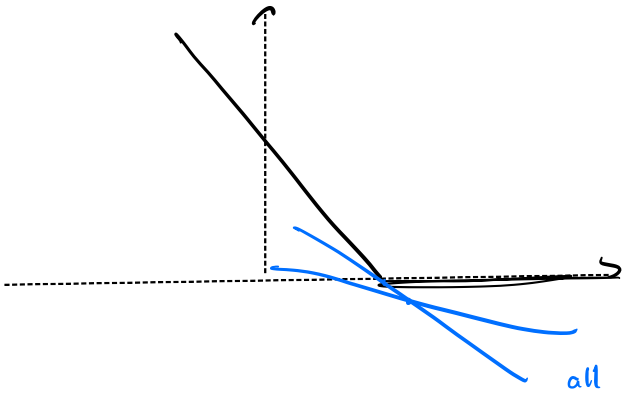
$\therefore \quad \varepsilon \cdot \|\nabla f(x)\| \leq \rho \varepsilon$.

By (3), Lipschitz condition & Lemma (iterative update), we are done.

Variants & other properties

* If f is not differentiable, we can use subgradient descent. A subgradient $\partial f(x)$ is any vector which satisfies first-order definition of convexity,

$$f(y) - f(x) \geq \langle y - x, \partial f(y) \rangle$$



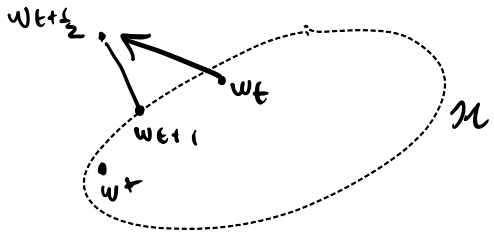
(provides lower bound on the function).

* If we are doing constrained optimization, we can use projected gradient descent.

Suppose we have: $\min_{x \in \mathcal{H}} f(x)$ ($f(x)$ convex, $\mathcal{H}$ convex)

$$w_{t+\frac{1}{2}} = w_t - \eta \, \nabla f(w_t)$$

$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{H}} \| w - w_{t+\frac{1}{2}} \|_2$$



Lemma: For any $v \in \mathcal{W}$

$$\|w_{t+1} - v\|_2 \leq \|w_{t+\frac{1}{2}} - v\|_2$$

Lemma Projected GD has the same convergence rate as GD.

* Some faster convergence guarantees for GD.

Def (smoothness) We say a function f is β-smooth if $\forall x, y \in \text{domain}(f)$,

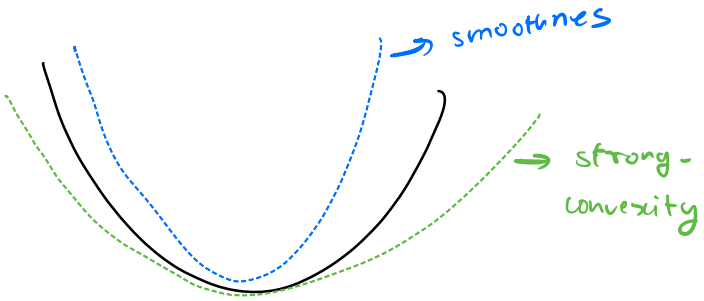
$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2.$$

This also implies that

$$f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\beta}{2} \|y - x\|_2^2.$$

Thm Let f be a convex function that is β-smooth. Then GD with step size $\eta = 1/\beta$ satisfies

$$f(w_T) - f(w^*) \leq \frac{B^2 \beta}{2T}.$$



Def (Strong convexity) A function f is λ-strongly convex if its domain is convex & $\forall x, y$

$$f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\lambda}{2} \|y - x\|^2.$$

Thm. Let f be a convex function that is β-smooth & λ-strongly convex. Let $\kappa = \beta/\lambda > 1$. Then GD with step size $\eta = \frac{2}{\lambda + \beta}$ satisfies

$$f(w_T) - f(w^*) \leq e^{-T/\kappa} \frac{\beta^2}{2}.$$

(To get error ε, only need $O\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$ samples).

E.g.,

$$f(x) = \min_{x \in \mathbb{R}^d} x^T A x$$

convex if $A \succeq 0$

κ is ratio of largest to smallest eigenvalue

Bonus There are accelerated methods which can faster (& optimal!) convergence rates.

Stochastic gradient descent (SGD)

1. Initialize w_1
2. For $t = 1, \ldots, T$
3. Choose any v_t s.t. $\mathbb{E}[v_t \mid w_t] = \nabla f(w_t)$.
4. $w_{t+1} \leftarrow w_t - \eta v_t$

If we are minimizing training loss over n points, just sample one point (x_i, y_i) uniformly at random & take gradient at (x_i, y_i).

$$\nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla_w \ell(w, z_i)$$