Exponential weights

UNIVERSITY of WASHINGTON

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	…
Expert 1	.78	.15				
Expert 2	.15	.25	- - - .			
Expert 3	.25	.38				

ith expert has loss $z_t(i) = |b_t(i) - y_t|$, $y_t \in [0,1]$ is truth

$I_t \sim p_t$

Choose $I_t \in [d] = \{1, \ldots, d\}$ at each time t

Regret $\max_i \sum_{t=1}^{T} z_t(I_t) - z_t(i) = \sum_{t=1}^{T} z_t(I_t) - \min_{i \in [d]} \sum_t z_t(i)$

Pseudo-Regret $\max_i \mathbb{E}\left[\sum_{t=1}^{T} z_t(I_t) - z_t(i)\right] = \underline{\max_i \sum_{t=1}^{T} \langle z_t, p_t \rangle - \langle z_t, e_i \rangle}$

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	…
Expert 1						
Expert 2						
Expert 3						

$$\Delta_d = \{ x \in \mathbb{R}^d : x_i \geq 0, \forall i, \sum_{i=1}^{d} x_i = 1 \}$$

$$z_t(i) = |b_t(i) - y_t|$$

ith expert's prediction

True temperature

$\texttt{Input:}$ d experts

$\texttt{for}$ $t = 1, 2, \ldots$

Player picks $\underline{p_t \in \Delta_d}$ and plays $I_t \sim p_t$

Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$

Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	…
Expert 1						
Expert 2						
Expert 3						

$$z_t(i) = |b_t(i) - y_t|$$

*i*th expert's prediction

True temperature

$\texttt{Input:}$ d experts
$\texttt{for}$ $t = 1, 2, \ldots$
 Player picks $p_t \in \triangle_d$ and plays $I_t \sim p_t$
 Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
 Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

Goal: Minimize regret wrt best $\max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle$

Expert prediction

Goal: Minimize regret wrt best $\max_{i\in[d]} \sum_{t=1}^{T} \langle p_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle$

`Input:` d experts
`for` $t = 1, 2, \ldots$
  Player picks $p_t \in \triangle_d$ and plays $I_t \sim p_t$
  Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
  Player pays loss $\langle p_t, z_t\rangle = \mathbb{E}[z_t(I_t)]$

Exponential weights algorithm

`Input:` d experts, $\eta > 0$
`Initialize:` $w_1 \in [1, \ldots, 1]^\top \in \mathbb{R}^d$
`for` $t = 1, 2, \ldots$
  Player plays $I_t \sim p_t$ where $p_t(i) = w_t(i)/\sum_{j=1}^d w_t(j)$
  Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
  Player pays loss $\langle p_t, z_t\rangle = \mathbb{E}[z_t(I_t)]$
  Player updates weights $w_{t+1}(i) = \underline{w_t(i)\exp(-\eta z_t(i))}$ $= \exp\left(-\eta \sum_{s=1}^{t} z_s(i)\right)$

Expert prediction

Goal: Minimize regret wrt best $\max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle$

Exponential weights algorithm

$\texttt{Input:}$ d experts, $\eta > 0$

$\texttt{Initialize:}$ $w_1 \in [1, \ldots, 1]^\top \in \mathbb{R}^d$

$\texttt{for}$ $t = 1, 2, \ldots$

Player plays $I_t \sim p_t$ where $p_t(i) = w_t(i) / \sum_{j=1}^{d} w_t(j)$

Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$

Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

Player updates weights $w_{t+1}(i) = w_t(i) \exp(-\eta z_t(i))$

Theorem: If $z_t \in [0,1]^d$ $\forall t$, and I_t, p_t are chosen by exponential weights then $\max_{i \in [d]} \mathbb{E}\left[\sum_{t=1}^{T} \langle I_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle\right] = \max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle \leq \frac{\log(d)}{\eta} + \frac{T\eta}{8}$

Choosing $\eta = \sqrt{\frac{8 \log(d)}{T}}$ gives regret bound of $\sqrt{T \log(d)/2}$

Expert prediction

Goal: Minimize regret wrt best $\max_{i\in[d]} \sum_{t=1}^{T} \langle p_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle$

$$\log AB = \log A + \log B$$

Exponential weights algorithm, proof: Let $W_t = \sum_{i=1}^{d} w_t(i)$ so that

$$\log \frac{W_{T+1}}{W_1} = \log \frac{\sum_i w_{T+1}(i)}{d} \geq \log\left(w_{T+1}(i)/d\right) = -\eta \sum_{t=1}^{T} z_t(i) - \log(d)$$

$$\log \frac{W_{T+1}}{W_1} = \sum_{t=1}^{T} \log\left(\frac{\sum_i w_{t+1}(i)}{\sum_j w_t(j)}\right) = \sum_t \log\left(\sum_i \frac{w_t(i)}{\sum_j w_t(j)} \cdot \exp(-\eta z_t(i))\right)$$

$\mu = -\langle z_t, p_t\rangle$

$$= \sum_t \log\left(\sum_i p_t(i) \exp(-\eta z_t(i))\right) = \sum_t \log\left(\mathbb{E}_{I_t\sim p_t}\left[\exp(-\eta z_t(I_t))\right]\right)$$

$$= \sum_t \log\left(e^{-\eta\langle z_t, p_t\rangle} \underbrace{\mathbb{E}\left[\exp\left(-\eta\left(z_t(I_t) - \langle z_t, p_t\rangle\right)\right)\right]}_{\leq \exp(\eta^2/8)}\right)$$

$$\leq \sum_t -\eta\langle z_t, p_t\rangle + T\eta^2/8$$

Expert prediction

Goal: Minimize regret wrt best $\max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle$

Exponential weights algorithm, proof: Let $W_t = \sum_{i=1}^{d} w_t(i)$ so that

$$
\begin{aligned}
\log \frac{W_{T+1}}{W_1} &= \sum_{t=1}^{T} \log \frac{W_{t+1}}{W_t} \\
&= \sum_{t=1}^{T} \log \left(\sum_{i=1}^{d} \frac{w_{t+1}(i)}{W_t} \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{i=1}^{d} \frac{w_t(i) \exp(-\eta z_t(i))}{W_t} \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{i=1}^{d} p_t(i) \exp(-\eta z_t(i)) \right) \\
&= \sum_{t=1}^{T} \log \left(\exp(-\eta \mathbb{E}[z_t(I_t)]) \sum_{i=1}^{d} p_t(i) \exp(-\eta (z_t(i) - \mathbb{E}[z_t(I_t)])) \right) \\
&= \sum_{t=1}^{T} -\eta \mathbb{E}[z_t(I_t)] + \log \left(\mathbb{E}[\exp(-\eta(z_t(I_t) - \mathbb{E}[z_t(I_t)]))] \right) \\
&\le \sum_{t=1}^{T} -\eta \mathbb{E}[z_t(I_t)] + \eta^2/8
\end{aligned}
$$

$$
\begin{aligned}
\log \frac{W_{T+1}}{W_1} &\ge \log \frac{w_{T+1}(i)}{W_1} \\
&= -\log(d) + \log \left(\prod_{t=1}^{T} \exp(-\eta z_t(i)) \right) \\
&= -\log(d) - \sum_{t=1}^{T} \eta z_t(i)
\end{aligned}
$$

Lemma (Hoeffding's Lemma). Let X be a real-valued random variable such that $X \in [a, b]$ almost surely, and let $\mathbb{E}[X] = \mu$. Then, for any $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{t(X-\mu)}\right] \le \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

$$\implies \sum_{t=1}^{T} \eta \mathbb{E}[z_t(I_t)] - \sum_{t=1}^{T} \eta z_t(i) \le \log(d) + \eta^2 T/8$$

Path-length bounds

There exists an algorithm $\left(\tilde{p}_t = (1 - \frac{1}{T}) p_t + \frac{1}{T} (\frac{1}{d} \mathbf{1}) \right)$

$$\sum_{t=1}^{T} \langle z_t, \tilde{p}_t \rangle - \sum_t \langle z_t, q_t \rangle \leq \sqrt{T \log d} + \sqrt{T \sum_{t=1}^{T} \| q_{t+1} - q_t \|_1}$$

for any $q_t \in [0,1]^d \ \forall t$

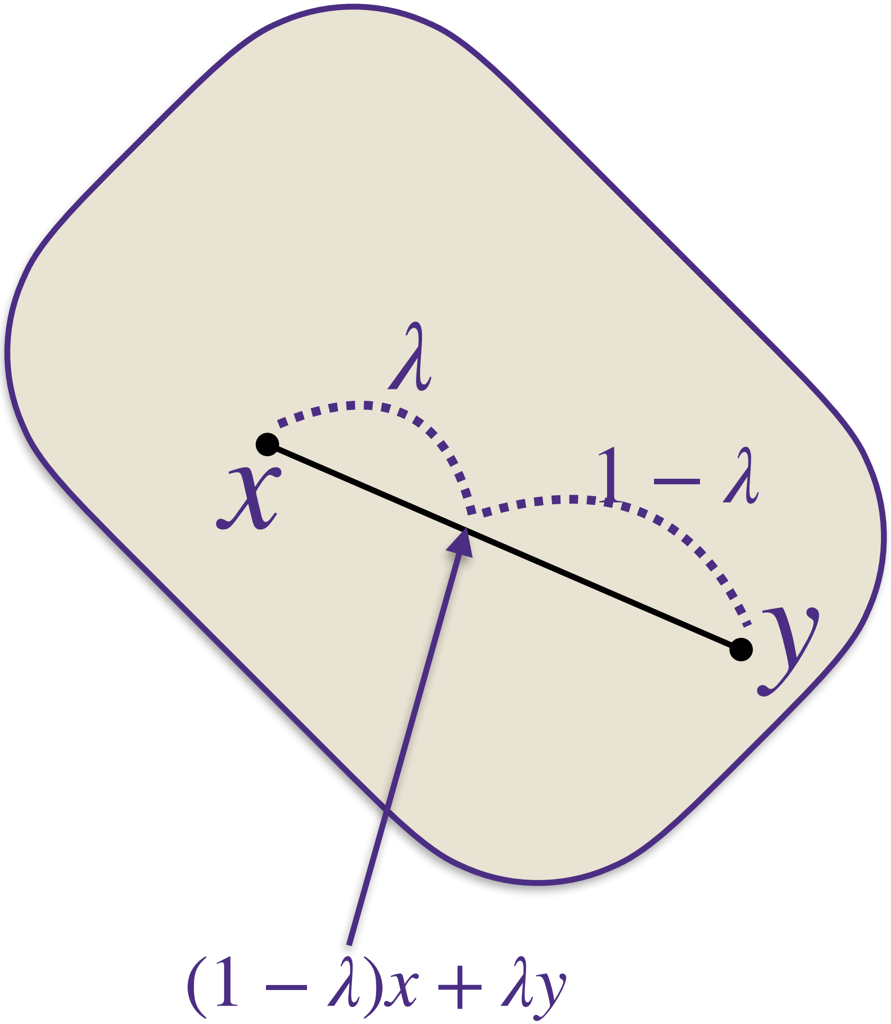
$$\sum_{t=1}^{T} \langle z_t, p_t \rangle - \min_i \sum_t \langle z_t, e_i \rangle \leq \sqrt{\min_i \sum_t z_t(i) \cdot \log(d)} +$$

Convexity

- When is an optimization (or learning) easy/fast to solve?

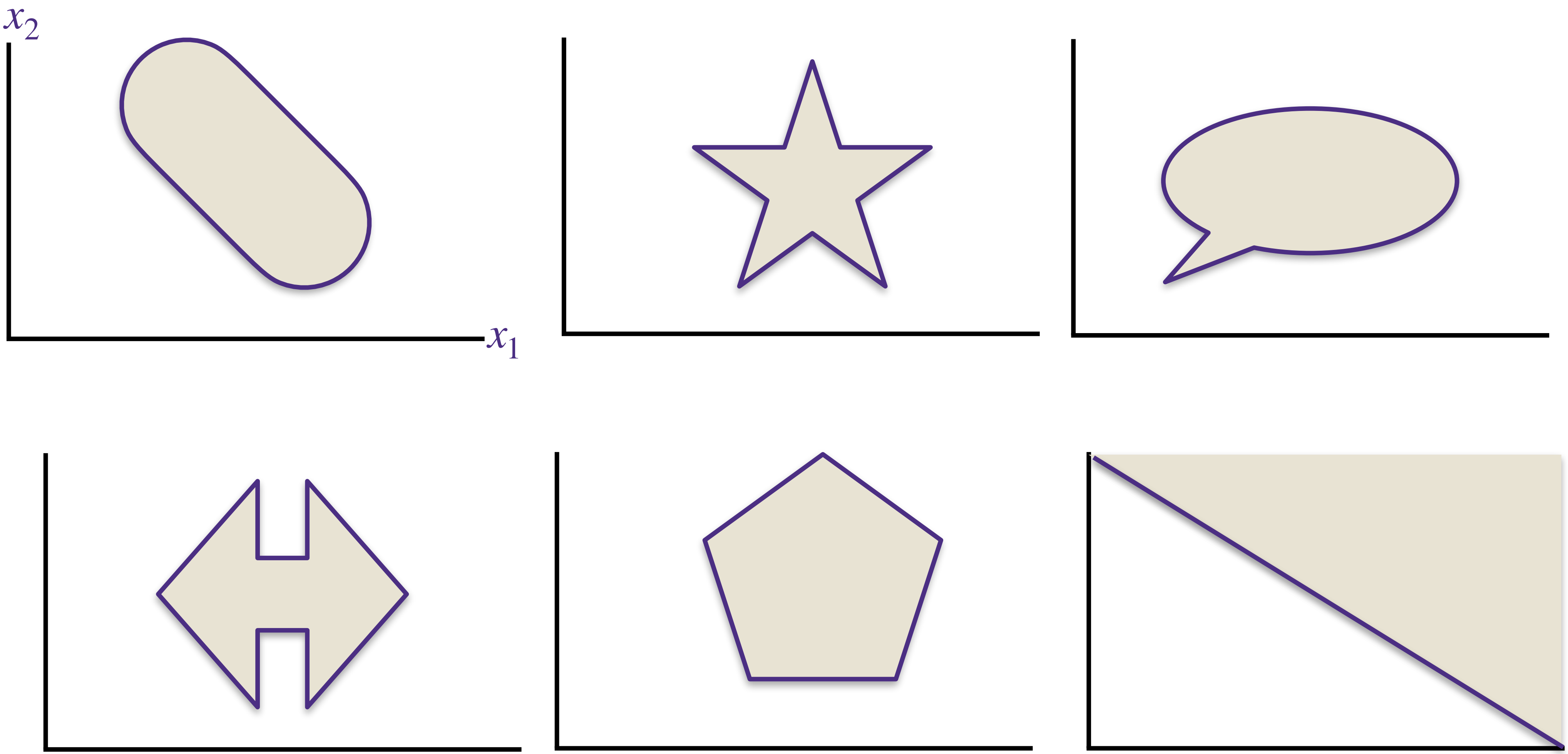
What is a convex set?

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$



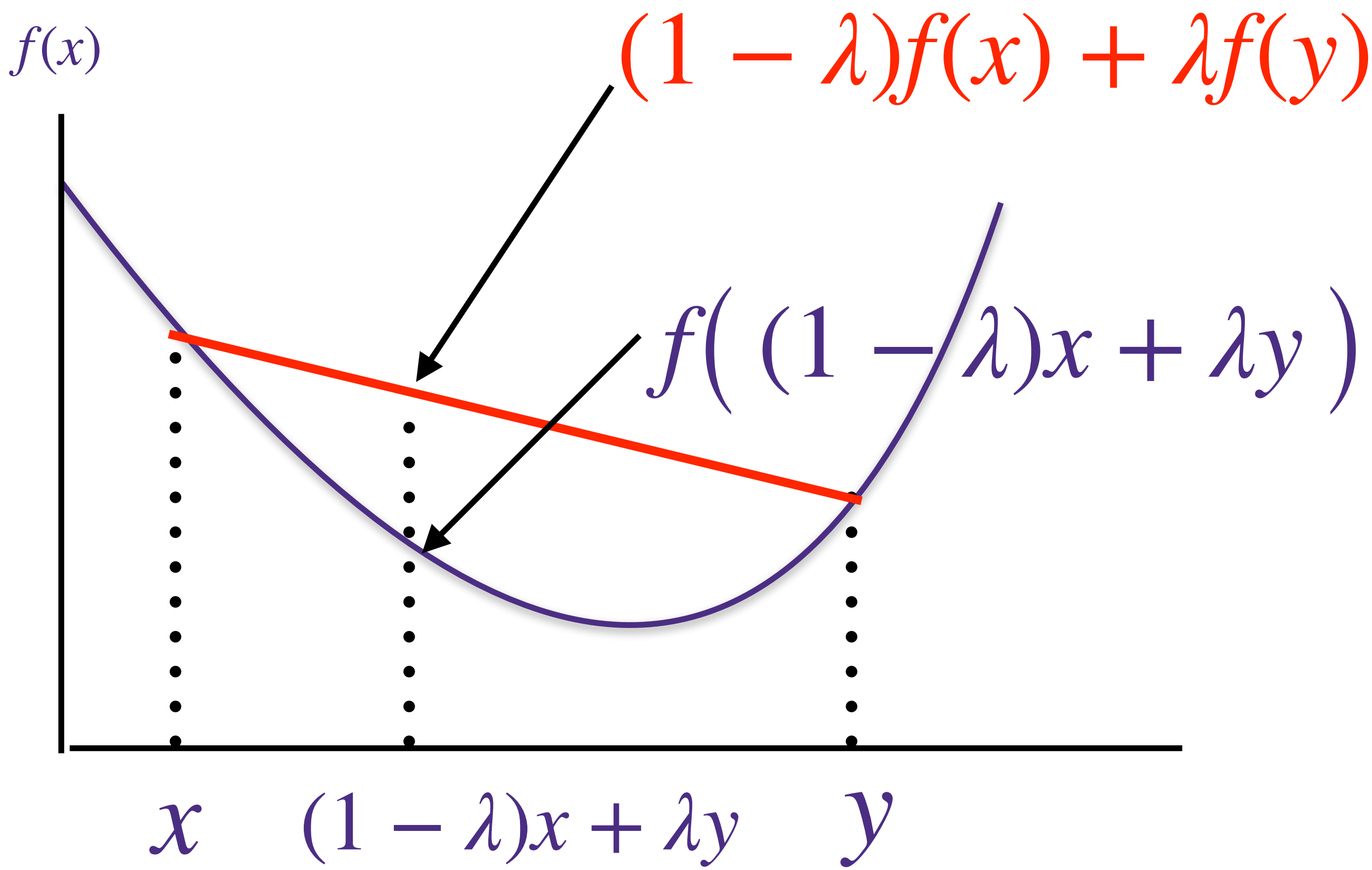
What is a convex set?

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$



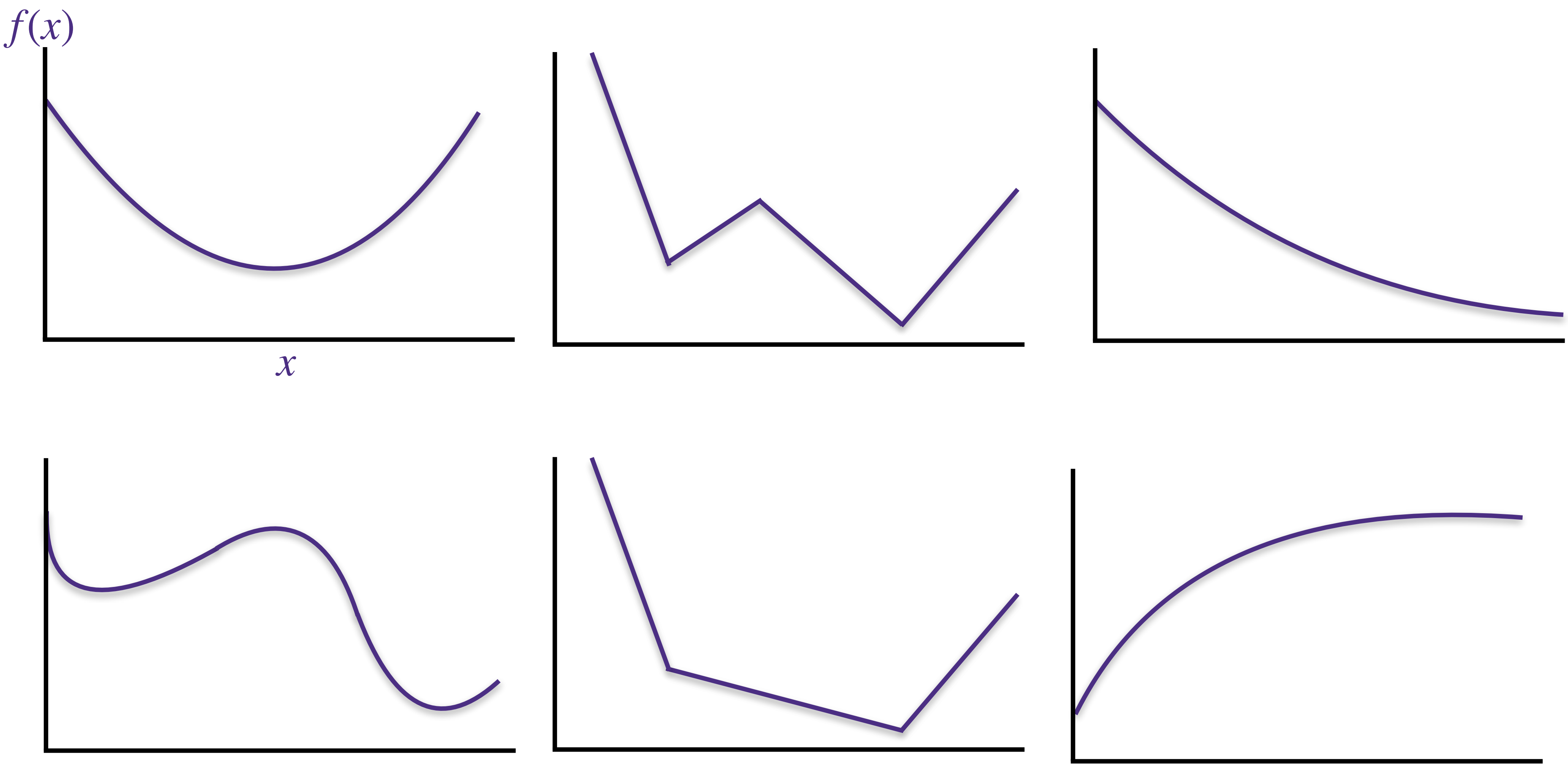
What is a convex function?

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$



What is a convex function?

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$



Convex functions and convex sets?

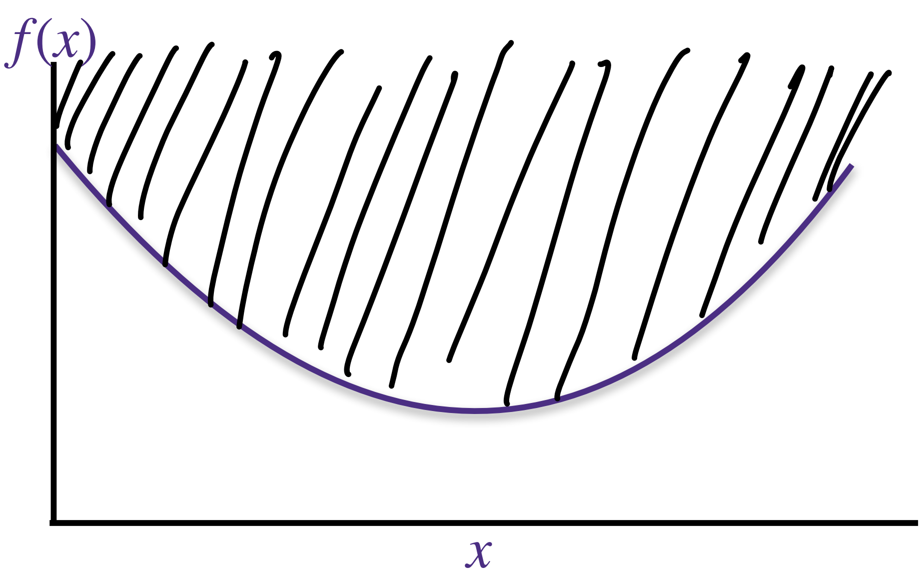
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

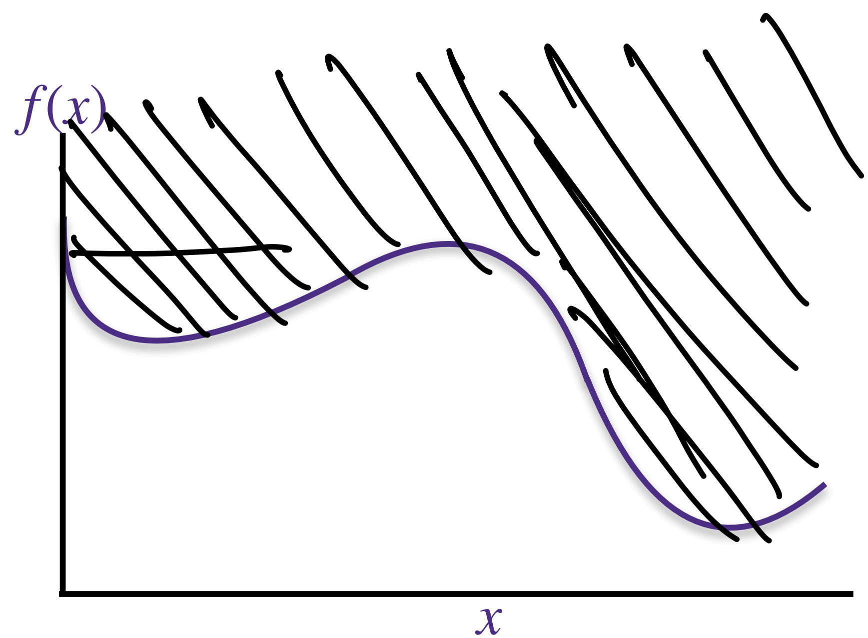
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

Graph of f id defined as $\{(x,t) \,:\, f(x) = t\}$

Epigraph of f is defined as $\{(x,t) \,:\, f(x) \leq t\}$



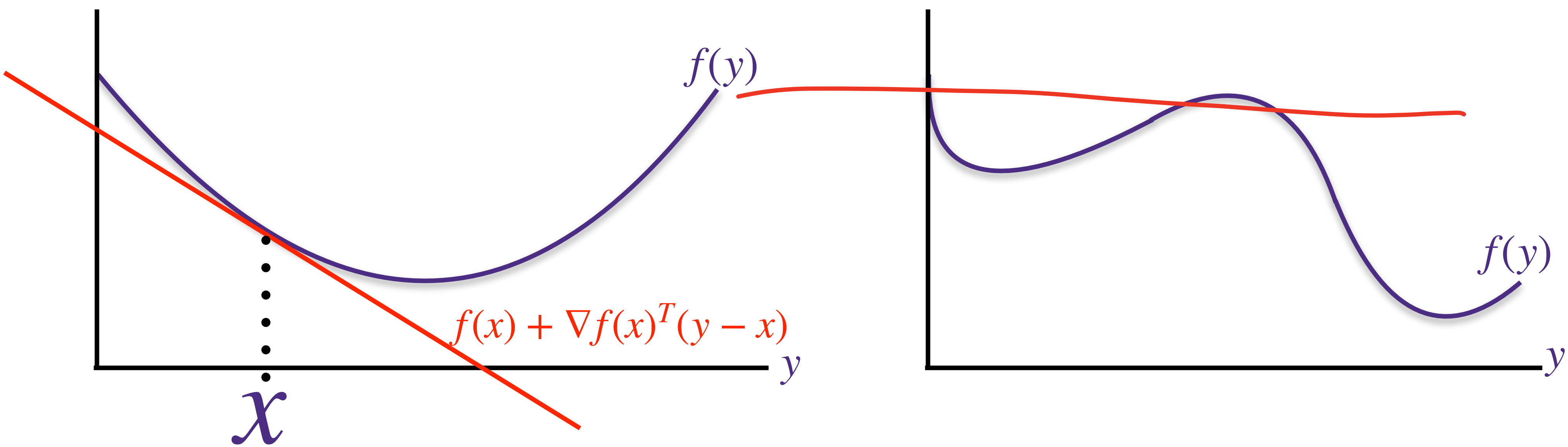


More definitions of convexity

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

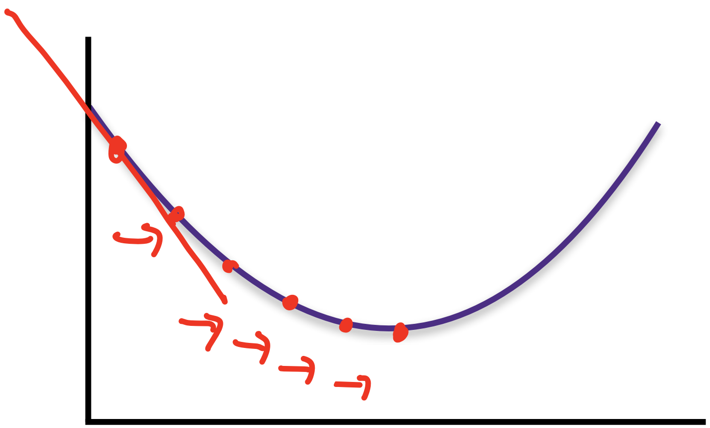


Why do we care about convexity?

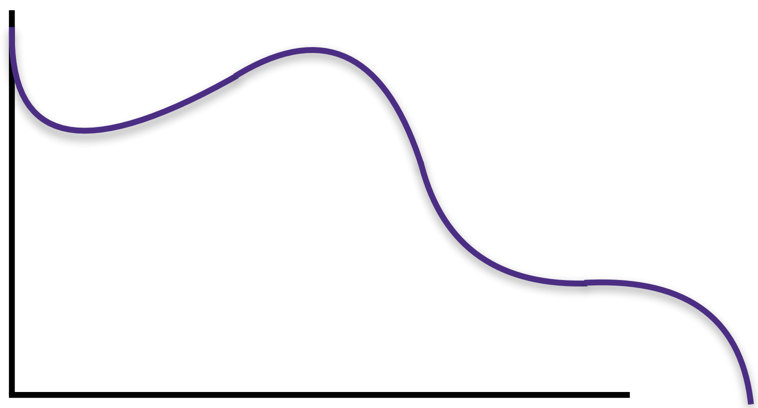
Convex functions

- All local minima are global minima
- Efficient to optimize (e.g., gradient descent)

Convex Function



Non-convex Function



We only need to find a point with $\nabla f(x) = 0$, which for convex functions implies that it is a local minima and a global minima

For non-convex functions, a stationary point with $\nabla f(x) = 0$ could be a local minima, a local maxima, or a saddle point

Online Convex Optimization

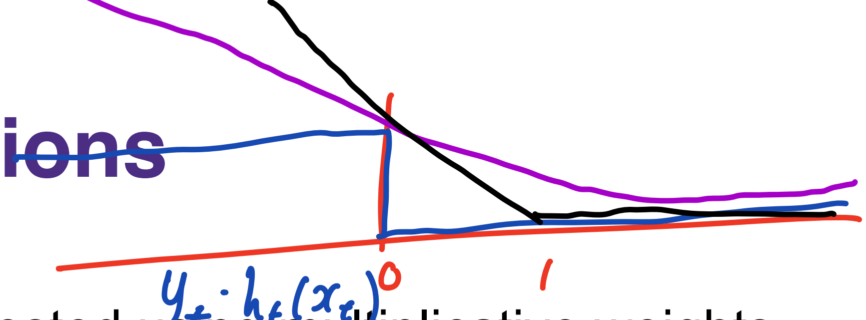
Convex surrogate loss functions

Previous section for the adversarial case suggested using multiplicative weights over the |H| hypotheses, which is completely intractable in practice.

And in the stochastic case we used $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$ which is also intractable to compute!

So it seems we have no practical algorithm! Solution: relax the objective.

Convex surrogate loss functions



Previous section for the adversarial case suggested using multiplicative weights over the |H| hypotheses, which is completely intractable in practice.

And in the stochastic case we used $h_t \in \arg\min_{h\in\mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$ which is also intractable to compute!

So it seems we have no practical algorithm! Solution: relax the objective.

Instead of $$\max_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

We use $$\max_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h_t,(x_t,y_t)) - \ell(h,(x_t,y_t))$$ with $\mathcal{H}$ convex

Example: Linear classification takes $\mathcal{H} \subset \mathbb{R}^d$ and $\ell(h,(x_t,y_t)) = \log(1+\exp(-y_t h^\top x_t))$

Convex surrogate loss functions

$h_t(x) := \langle h, x \rangle$

$h \in B_2$

Goal: $\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h_t, (x_t, y_t)) - \ell(h, (x_t, y_t))$ with $\mathcal{H}$ convex

Online gradient descent

Input: $\mathcal{H} \subset \mathbb{R}^d$, convex loss function ℓ, step size $\eta > 0$

Initialize: Choose any $h_1 \in \mathcal{H}$

for $t = 1, 2, \ldots$

Player plays $h_t \in \mathcal{H}$

Adversary simultaneously reveals (x_t, y_t)

Player pays loss $\ell_t(h_t) := \ell(h_t, (x_t, y_t)) = \max\{0, 1 - y_t \langle x_t, h_t \rangle\}$

Player updates $h_{t+1} = \Pi_{\mathcal{H}}(h_t - \eta \nabla_h \ell_t(h_t))$

Theorem Online gradient descent satisfies for any $h_* \in \mathcal{H}$

$$\sum_{t=1}^{T} \ell(h_t, (x_t, y_t)) - \ell(h_*, (x_t, y_t)) \leq \frac{\|h_*\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_h \ell_t(h_t)\|_2^2$$

if $\max_{h \in \mathcal{H}} \|h_*\|_2 \leq R$ and $\ell(\cdot)$ is G-Lipschitz then regret$\leq RB\sqrt{T}$

Proof

Theorem Online gradient descent satisfies for any $h_* \in \mathcal{H}$

$$\sum_{t=1}^T \ell(h_t,(x_t,y_t)) - \ell(h_*,(x_t,y_t)) \le \frac{\|h_*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\nabla_h \ell_t(h_t)\|_2^2$$

$$\|h_{t+1} - h_*\|_2^2 = \|\Pi_{\mathcal{H}}(h_t - \eta\nabla\ell(h_t)) - \Pi_{\mathcal{H}}(h_*)\|_2^2$$

$$\le \|h_t - \eta\nabla\ell(h_t) - h_*\|_2^2$$

$$= \|h_t - h_*\|_2^2 - 2\eta\underbrace{\langle \nabla\ell(h_t), h_t - h_*\rangle}_{\ge \ell(h_t) - \ell(h_*))} + \eta^2\|\nabla\ell(h_t)\|_2^2$$

$$\le \|h_t - h_*\|_2^2 - 2\eta(\ell(h_t) - \ell(h_*)) + \eta^2\|\nabla\ell(h_t)\|_2^2$$

$$\|h_{T+1} - h_*\|_2^2 - \|h_1 - h_*\|_2^2 \le \sum_{t=1}^T -2\eta(\ \cdots\ ) + \eta^2\|\ \cdots\ \|_2^2$$

Proof

Theorem Online gradient descent satisfies for any $h_* \in \mathcal{H}$
$\sum_{t=1}^T \ell(h_t,(x_t,y_t)) - \ell(h_*,(x_t,y_t)) \leq \frac{\|h_*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\nabla_h \ell_t(h_t)\|_2^2$

$$
\begin{aligned}
\|h_{t+1} - h_*\|_2^2 &= \|\Pi_{\mathcal{H}}(h_{t+1}) - \Pi_{\mathcal{H}}(h_*)\|_2^2 \\
&= \|\Pi_{\mathcal{H}}(h_t - \eta \nabla \ell_t(h_t)) - \Pi_{\mathcal{H}}(h_*)\|_2^2 \\
&\leq \|h_t - \eta \nabla \ell_t(h_t) - h_*\|_2^2 \\
&= \|h_t - h_*\|_2^2 - 2\eta \nabla \ell_t(h_t)^\top (h_t - h_*) + \eta^2 \|\nabla \ell_t(h_t)\|_2^2 \\
&\leq \|h_t - h_*\|_2^2 - 2\eta\big(\ell_t(h_t) - \ell_t(h_*)\big) + \eta^2 \|\nabla \ell_t(h_t)\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{t=1}^T \big(\ell_t(h_t) - \ell_t(h_*)\big) &\leq \sum_{t=1}^T \frac{\|h_t - h_*\|_2^2 - \|h_{t+1} - h_*\|_2^2}{2\eta} + \sum_{t=1}^T \frac{\eta}{2}\|\nabla \ell_t(h_t)\|_2^2 \\
&\leq \frac{\|h_1 - h_*\|_2^2}{2\eta} + \sum_{t=1}^T \frac{\eta}{2}\|\nabla \ell_t(h_t)\|_2^2
\end{aligned}
$$