Online Convex Optimization, Online and Stochastic Gradient Descent

Dimitris Fotakis

School of Electrical and Computer Engineering
National Technical University of Athens, Greece

Online Convex Optimization

General framework: convex set $S \subseteq \mathbb{R}^d$

On each day $t = 1, \dots, T$:

1. Learner picks vector $p_t \in S$
2. Adversary picks **convex loss** function $f_t : S \to \mathbb{R}$, with f_t differentiable and L-Lipschitz wrt some norm $\| \cdot \|$, i.e., $|f_t(p) - f_t(p')| \leq L \cdot \|p - p'\|$
3. Learner **learns** f_t and incurs loss $f_t(p_t)$

Online Convex Optimization

General framework: convex set $S \subseteq \mathbb{R}^d$

On each day $t = 1, \ldots, T$:

1. Learner picks vector $p_t \in S$
2. Adversary picks **convex loss** function $f_t : S \to \mathbb{R}$, with f_t differentiable and L-Lipschitz wrt some norm $\|\cdot\|$, i.e., $|f_t(p) - f_t(p')| \le L \cdot \|p - p'\|$
3. Learner **learns f_t** and incurs loss $f_t(p_t)$

Goal is to minimize **regret**:

$$\text{Regret}(T) = \sup_{f_1,\ldots,f_T} \left(\sum_{t=1}^{T} f_t(p_t) - \min_{p \in S} \sum_{t=1}^{T} f_t(p) \right)$$

(Online) algorithm is **no-regret** if $\text{Regret}(T)/T \to 0$ at $T \to \infty$

Online Convex Optimization

General framework: convex set $S \subseteq \mathbb{R}^d$

On each day $t = 1, \ldots, T$:

1. Learner picks vector $p_t \in S$
2. Adversary picks **convex loss** function $f_t : S \to \mathbb{R}$, with f_t differentiable and L-Lipschitz wrt some norm $\|\cdot\|$, i.e., $|f_t(p) - f_t(p')| \leq L \cdot \|p - p'\|$
3. Learner **learns f_t** and incurs loss $f_t(p_t)$

Goal is to minimize **regret**:

$$\text{Regret}(T) = \sup_{f_1,\ldots,f_T} \left(\sum_{t=1}^{T} f_t(p_t) - \min_{p \in S} \sum_{t=1}^{T} f_t(p) \right)$$

(Online) algorithm is **no-regret** if $\text{Regret}(T)/T \to 0$ at $T \to \infty$

- **Experts**: $S = \Delta_d$, $f_t(p) = \langle p, \ell_t \rangle$ (linear expected loss)

Online Convex Optimization

General framework: convex set $S \subseteq \mathbb{R}^d$

On each day $t = 1, \ldots, T$:

1. Learner picks vector $p_t \in S$
2. Adversary picks **convex loss** function $f_t : S \to \mathbb{R}$, with f_t differentiable and L-Lipschitz wrt some norm $\|\cdot\|$, i.e., $|f_t(p) - f_t(p')| \leq L \cdot \|p - p'\|$
3. Learner **learns f_t** and incurs loss $f_t(p_t)$

Goal is to minimize **regret**:

$$\mathrm{Regret}(T) = \sup_{f_1,\ldots,f_T} \left(\sum_{t=1}^{T} f_t(p_t) - \min_{p \in S} \sum_{t=1}^{T} f_t(p) \right)$$

(Online) algorithm is **no-regret** if $\mathrm{Regret}(T)/T \to 0$ at $T \to \infty$

- **Experts**: $S = \Delta_d$, $f_t(p) = \langle p, \ell_t \rangle$ (linear expected loss)
- **Online Quadratic Optimization**: learner $p_t \in S$, adversary $z_t \in S$, $f_t(p) = \|p - z_t\|_2^2$.

Online Convex Optimization

General framework: convex set $S \subseteq \mathbb{R}^d$

On each day $t = 1, \ldots, T$:

1. Learner picks vector $p_t \in S$
2. Adversary picks **convex loss** function $f_t : S \to \mathbb{R}$, with f_t differentiable and L-Lipschitz wrt some norm $\|\cdot\|$, i.e., $|f_t(p) - f_t(p')| \leq L \cdot \|p - p'\|$
3. Learner **learns f_t** and incurs loss $f_t(p_t)$

Goal is to minimize **regret**:

$$\text{Regret}(T) = \sup_{f_1,\ldots,f_T} \left(\sum_{t=1}^{T} f_t(p_t) - \min_{p \in S} \sum_{t=1}^{T} f_t(p) \right)$$

(Online) algorithm is **no-regret** if $\text{Regret}(T)/T \to 0$ at $T \to \infty$

- **Experts**: $S = \Delta_d, f_t(p) = \langle p, \ell_t \rangle$ (linear expected loss)
- **Online Quadratic Optimization**: learner $p_t \in S$, adversary $z_t \in S, f_t(p) = \|p - z_t\|_2^2$.
- **Online Least Squares Linear Regression**: learner $p_t \in \mathbb{R}^d$, $\|p_t\| \leq B$, adversary $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}, f_t(p) = (\langle p, x_t \rangle - y_t)^2$

Follow / Be the Regularized Leader

$F_t(p) = \sum_{\tau=1}^{t} f_\tau(p)$ and $\tilde{F}_t(p) = \sum_{\tau=1}^{t} f_\tau(p) + R(p)/\eta$

FTRL: $\tilde{p}_t = \arg\min_{p \in S} \tilde{F}_{t-1}(p)$

BTRL: $\tilde{p}_t^* = \arg\min_{p \in S} \tilde{F}_t(p)$

Follow / Be the Regularized Leader

$F_t(p) = \sum_{\tau=1}^{t} f_\tau(p)$ and $\tilde{F}_t(p) = \sum_{\tau=1}^{t} f_\tau(p) + R(p)/\eta$

FTRL: $\tilde{p}_t = \arg\min_{p \in S} \tilde{F}_{t-1}(p)$

BTRL: $\tilde{p}_t^* = \arg\min_{p \in S} \tilde{F}_t(p)$

$1/\eta$-**strongly convex** function $R : S \to \mathbb{R}$ wrt norm $\|\cdot\|$, if $\forall\, x, y \in S$:

$$R(x) \geq R(y) + \langle \nabla R(y), x - y \rangle + \frac{1}{2\eta}\|x - y\|^2$$

Follow / Be the Regularized Leader

$F_t(p) = \sum_{\tau=1}^{t} f_\tau(p)$ and $\tilde{F}_t(p) = \sum_{\tau=1}^{t} f_\tau(p) + R(p)/\eta$

FTRL: $\tilde{p}_t = \arg\min_{p \in S} \tilde{F}_{t-1}(p)$

BTRL: $\tilde{p}_t^* = \arg\min_{p \in S} \tilde{F}_t(p)$

$1/\eta$-**strongly convex** function $R : S \to \mathbb{R}$ wrt norm $\|\cdot\|$, if $\forall\, x, y \in S$:

$$R(x) \geq R(y) + \langle \nabla R(y), x - y \rangle + \frac{1}{2\eta}\|x - y\|^2$$

Functions $f, g : S \to \mathbb{R}$ be $1/\eta$-**strongly convex** wrt some norm $\|\cdot\|$ and $h(x) = g(x) - f(x)$ be L-**Lipschitz** wrt $\|\cdot\|$.
Then, $\|x_f^* - x_g^*\| \leq \eta \cdot L$, with x_f^*, x_g^* **minimizers** of f, g.

Regret of FTRL Against BTRL

$$\begin{aligned}\text{Regret}_{FTRL}(T) &\leq \text{Regret}_{BTRL}(T) + L \cdot \sum_{t=1}^{T} \|\tilde{p}_t - \tilde{p}_{t+1}\| \\ &\leq \text{Regret}_{BTRL}(T) + \eta \cdot L^2 \cdot T\end{aligned}$$

Regret of FTRL Against BTRL

$$\mathrm{Regret}_{FTRL}(T) \leq \mathrm{Regret}_{BTRL}(T) + L \cdot \sum_{t=1}^{T} \|\tilde{p}_t - \tilde{p}_{t+1}\|$$
$$\leq \mathrm{Regret}_{BTRL}(T) + \eta \cdot L^2 \cdot T$$

Proof : Second inequality from strong convexity, because $\tilde{p}_t, \tilde{p}_{t+1}$ are minimizers of $1/\eta$-strong convex functions $\tilde{F}_{t-1}(p)$ and $\tilde{F}_t(p)$ with difference $f_t(p)$ which is L-Lipschitz.

Regret of FTRL Against BTRL

$$\begin{aligned}\text{Regret}_{FTRL}(T) &\le \text{Regret}_{BTRL}(T) + L \cdot \sum_{t=1}^{T} \|\tilde{p}_t - \tilde{p}_{t+1}\| \\ &\le \text{Regret}_{BTRL}(T) + \eta \cdot L^2 \cdot T\end{aligned}$$

Proof : Second inequality from strong convexity, because $\tilde{p}_t, \tilde{p}_{t+1}$ are minimizers of $1/\eta$-strong convex functions $\tilde{F}_{t-1}(p)$ and $\tilde{F}_t(p)$ with difference $f_t(p)$ which is L-Lipschitz.

For the first inequality, we observe that:

$$\begin{aligned}\text{Regret}_{FTRL}(T) - \text{Regret}_{BTRL}(T) &= \sum_{t=1}^{T} (f_t(\tilde{p}_t) - f_t(\tilde{p}_t^*)) \\ &\le L \sum_{t=1}^{T} \|\tilde{p}_t - \tilde{p}_t^*\| \\ &= L \sum_{t=1}^{T} \|\tilde{p}_t - \tilde{p}_{t+1}\|\end{aligned}$$

Regret of Be the Regularized Leader

$$\text{Regret}_{BTRL}(T) \leq \frac{1}{\eta}\left(\max_{p \in S} R(p) - \min_{p \in S} R(p)\right)$$

Regret of Be the Regularized Leader

$$\text{Regret}_{BTRL}(T) \leq \frac{1}{\eta}\left(\max_{p\in S} R(p) - \min_{p\in S} R(p)\right)$$

Proof :

- Let $f_0(p) = R(p)/\eta$ and $\tilde{p}_0^* = \arg\min_{p\in S} R(p)/\eta$.
- Using induction on t, we show that for all $t \geq 0$,

$$\sum_{\tau=0}^{t} f_\tau(\tilde{p}_\tau^*) \leq \min_{p\in S} \sum_{\tau=0}^{t} f_\tau(p) \quad \text{(notice fake action } \tilde{p}_0^* \text{ at } \tau = 0\text{)}$$

Regret of Be the Regularized Leader

$$\text{Regret}_{BTRL}(T) \leq \frac{1}{\eta}\left(\max_{p\in S} R(p) - \min_{p\in S} R(p)\right)$$

Proof:

- Let $f_0(p) = R(p)/\eta$ and $\tilde{p}_0^* = \arg\min_{p\in S} R(p)/\eta$.
- Using induction on t, we show that for all $t \geq 0$,

$$\sum_{\tau=0}^{t} f_\tau(\tilde{p}_\tau^*) \leq \min_{p\in S} \sum_{\tau=0}^{t} f_\tau(p) \quad \text{(notice fake action } \tilde{p}_0^* \text{ at } \tau = 0\text{)}$$

- Then, using the claim above,

$$\sum_{t=0}^{T} f_t(\tilde{p}_t^*) \leq \min_{p\in S} \sum_{t=0}^{T} f_t(p) \leq \max_{p\in S} f_0(p) + \min_{p\in S} \sum_{t=1}^{T} f_t(p)$$

Regret of Be the Regularized Leader

$$\text{Regret}_{BTRL}(T) \leq \frac{1}{\eta}\left(\max_{p\in S} R(p) - \min_{p\in S} R(p)\right)$$

Proof :

- Let $f_0(p) = R(p)/\eta$ and $\tilde{p}_0^* = \arg\min_{p\in S} R(p)/\eta$.
- Using induction on t, we show that for all $t \geq 0$,

$$\sum_{\tau=0}^{t} f_\tau(\tilde{p}_\tau^*) \leq \min_{p\in S} \sum_{\tau=0}^{t} f_\tau(p) \quad \text{(notice fake action } \tilde{p}_0^* \text{ at } \tau = 0\text{)}$$

- Then, using the claim above,

$$\sum_{t=0}^{T} f_t(\tilde{p}_t^*) \leq \min_{p\in S} \sum_{t=0}^{T} f_t(p) \leq \max_{p\in S} f_0(p) + \min_{p\in S} \sum_{t=1}^{T} f_t(p)$$

- Hence, by rearranging:

$$\sum_{t=1}^{T} f_t(\tilde{p}_t^*) - \min_{p\in S} \sum_{t=1}^{T} f_t(p) \leq \max_{p\in S} R(p)/\eta - \min_{p\in S} R(p)/\eta$$

Regret of Follow the Regularized Leader

Theorem :

$$\text{Regret}_{FTRL}(T) \leq \eta \cdot L^2 \cdot T + \frac{\left(\max_{p \in S} R(p) - \min_{p \in S} R(p)\right)}{\eta}$$

Regret of Follow the Regularized Leader

Theorem :

$$\text{Regret}_{FTRL}(T) \leq \eta \cdot L^2 \cdot T + \frac{\left(\max_{p \in S} R(p) - \min_{p \in S} R(p)\right)}{\eta}$$

Let $R^* = \max_{p \in S} R(p) - \min_{p \in S} R(p)$.

Setting $\eta = \sqrt{R^*/T}$ yields $\text{Regret}_{FTRL}(T) \leq (L^2 + 1)\sqrt{R^* T}$

Regret of Follow the Regularized Leader

Theorem :

$$\text{Regret}_{FTRL}(T) \le \eta \cdot L^2 \cdot T + \frac{\left(\max_{p\in S} R(p) - \min_{p\in S} R(p)\right)}{\eta}$$

Let $R^* = \max_{p\in S} R(p) - \min_{p\in S} R(p)$.

Setting $\eta = \sqrt{R^*/T}$ yields $\text{Regret}_{FTRL}(T) \le (L^2+1)\sqrt{R^* T}$

Multiplicative Weight Updates :

- Negative entropy $E^-(p) = \sum_{i=1}^{d} p_i \ln(p_i)$ is 1-strongly convex wrt L_1 norm.
- Using $E^-(p)$ as regularizer, results in the following update rule for linear losses $f_t(p) = \langle p, \ell_t \rangle$:

$$p_{t+1}(i) = p_t(i) \cdot e^{-\eta \ell_t(i)} \approx p_t(i)(1 - \eta \ell_t(i))$$

Regret of Follow the Regularized Leader

Theorem:

$$\text{Regret}_{FTRL}(T) \le \eta \cdot L^2 \cdot T + \frac{\left(\max_{p \in S} R(p) - \min_{p \in S} R(p)\right)}{\eta}$$

Let $R^* = \max_{p \in S} R(p) - \min_{p \in S} R(p)$.

Setting $\eta = \sqrt{R^*/T}$ yields $\text{Regret}_{FTRL}(T) \le (L^2 + 1)\sqrt{R^* T}$

Multiplicative Weight Updates:

- Negative entropy $E^-(p) = \sum_{i=1}^d p_i \ln(p_i)$ is 1-strongly convex wrt L_1 norm.
- Using $E^-(p)$ as regularizer, results in the following update rule for linear losses $f_t(p) = \langle p, \ell_t \rangle$:

$$p_{t+1}(i) = p_t(i) \cdot e^{-\eta \ell_t(i)} \approx p_t(i)(1 - \eta \ell_t(i))$$

- If $\ell_t \in [0,1]^d$, setting $\eta = \sqrt{\ln(d)/T}$, yields regret $2\sqrt{T \ln(d)}$

Online Gradient Descent

Online Projected Gradient Descent

Input: convex set S, $\boldsymbol{w}_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play $\boldsymbol{w}_t$, get **convex cost** function $f_t : S \to \mathbb{R}$, incur cost $f_t(\boldsymbol{w}_t)$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{w \in S} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Online Gradient Descent

Online Projected Gradient Descent

Input: convex set S, $\boldsymbol{w}_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play $\boldsymbol{w}_t$, get **convex cost** function $f_t : S \to \mathbb{R}$, incur cost $f_t(\boldsymbol{w}_t)$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Ignore project step for the analysis: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)$
Also we let $\boldsymbol{v}_t = \nabla f_t(\boldsymbol{w}_t)$ for brevity.

Online Gradient Descent

Online Projected Gradient Descent

Input: convex set S, $w_1 \in S$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Play w_t, get **convex cost** function $f_t : S \to \mathbb{R}$, incur cost $f_t(w_t)$
- Update $y_{t+1} = w_t - \eta \nabla f_t(w_t)$
- Project $w_{t+1} = \arg\min_{w \in S} \|w - y_{t+1}\|$

Ignore project step for the analysis: $w_{t+1} = w_t - \eta \nabla f_t(w_t)$
Also we let $v_t = \nabla f_t(w_t)$ for brevity.

Using convexity $f_t(w^*) \geq f(w_t) + \nabla f_t(w_t)(w^* - w_t)$, we get that:

$$\text{Regret}_{OGD} = \sum_{t=1}^{T} (f_t(w_t) - f_t(w^*)) \leq \sum_{t=1}^{T} v_t(w_t - w^*) \tag{1}$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2\|\boldsymbol{v}_t\|^2 \quad \Rightarrow\end{aligned}$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned} \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow \end{aligned}$$

$$\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) = \tfrac{1}{2\eta}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \qquad (2)$$

Online Gradient Descent

From the update rule, we get that:

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{v}_t - \boldsymbol{w}^*\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow\end{aligned}$$

$$\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) = \tfrac{1}{2\eta}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \qquad (2)$$

Substituting (2) in (1) results in a telescopic sum. So, we get that:

$$\begin{aligned}\text{Regret}_{OGD} = \sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) &\le \frac{\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2 \\ &\le \frac{B^2}{2\eta} + \frac{\eta T G^2}{2} \overset{\eta = \frac{B}{G\sqrt{T}}}{=} BG\sqrt{T}\end{aligned}$$

Similar regret with step $\eta_t = \frac{B}{G\sqrt{t}}$, because $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \le 2\sqrt{T}$.

α-Strongly Convex Functions

Using α-strong convexity

$$f_t(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_t) + \nabla f_t(\boldsymbol{w}_t)(\boldsymbol{w}^* - \boldsymbol{w}_t) + \frac{\alpha}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_t\|^2 \,,$$

we get that:

$$2 \sum_{t=1}^{T} (f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq \sum_{t=1}^{T} \left(2\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) - \alpha \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 \right) \quad (3)$$

α-Strongly Convex Functions

Using α-strong convexity

$$f_t(\boldsymbol{w}^*) \geq f(\boldsymbol{w}_t) + \nabla f_t(\boldsymbol{w}_t)(\boldsymbol{w}^* - \boldsymbol{w}_t) + \tfrac{\alpha}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_t\|^2 \,,$$

we get that:

$$2\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq \sum_{t=1}^{T}\left(2\boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*) - \alpha\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\right) \quad (3)$$

Substituting (2) in (3), we get that:

$$\begin{aligned} 2\text{Regret}_{OGD} &= 2\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \\ &\leq \sum_{t=1}^{T}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha\right) + \sum_{t=1}^{T}\eta_t\|\boldsymbol{v}_t\|^2 \\ &\overset{\eta_t = 1/(\alpha t)}{\leq} 0 + \frac{G^2(1 + \ln T)}{\alpha} \end{aligned}$$

Learning Using Stochastic Gradient Descent

- In learning problems, we want to solve $\min_{w \in H} L_{\mathcal{D}}(\boldsymbol{w})$, where $L_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)]$ and $\ell(\boldsymbol{w}, z)$ is the loss on z under $\boldsymbol{w} \in H$.

Learning Using Stochastic Gradient Descent

- In learning problems, we want to solve $\min_{w \in H} L_{\mathcal{D}}(\boldsymbol{w})$, where $L_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)]$ and $\ell(\boldsymbol{w}, z)$ is the loss on z under $\boldsymbol{w} \in H$.
- Instead of **Empirical Risk Minimization** $\min_{w \in H} L_S(\boldsymbol{w})$, we try to minimize $L_{\mathcal{D}}(\boldsymbol{w})$ directly using SGD and the fact that:

$$\nabla L_{\mathcal{D}}(\boldsymbol{w}) = \nabla \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)] = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\nabla \ell(\boldsymbol{w}, z)]$$

Learning Using Stochastic Gradient Descent

- In learning problems, we want to solve $\min_{w \in H} L_{\mathcal{D}}(\boldsymbol{w})$, where $L_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)]$ and $\ell(\boldsymbol{w}, z)$ is the loss on z under $\boldsymbol{w} \in H$.
- Instead of **Empirical Risk Minimization** $\min_{w \in H} L_S(\boldsymbol{w})$, we try to minimize $L_{\mathcal{D}}(\boldsymbol{w})$ directly using SGD and the fact that:

$$\nabla L_{\mathcal{D}}(\boldsymbol{w}) = \nabla \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\ell(\boldsymbol{w}, z)] = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\nabla \ell(\boldsymbol{w}, z)]$$

- We cannot calculate $\nabla L_{\mathcal{D}}(\boldsymbol{w})$, because we do not know $\mathcal{D}$.
- But expected value of $\nabla \ell(\boldsymbol{w}, z)$, over $z \sim \mathcal{D}$, is $\nabla L_{\mathcal{D}}(\boldsymbol{w})$
- We use $\nabla \ell(\boldsymbol{w}, z)$, for sample $z \sim \mathcal{D}$, which is an **unbiased estimator** of the gradient.

Stochastic Gradient Descent

Stochastic Gradient Descent

Input: convex set H, $\boldsymbol{w}_1 \in H$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Random sample $z_t \sim \mathcal{D}$, $\boldsymbol{v}_t = \nabla \ell(\boldsymbol{w}_t, z_t)$ and $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{v}_t$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{w \in H} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Return $\bar{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$

Stochastic Gradient Descent

Stochastic Gradient Descent

Input: convex set H, $\boldsymbol{w}_1 \in H$, time horizon T, step size η

For each $t = 1, \ldots, T$ do:

- Random sample $z_t \sim \mathcal{D}$, $\boldsymbol{v}_t = \nabla \ell(\boldsymbol{w}_t, z_t)$ and $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$
- Update $\boldsymbol{y}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{v}_t$
- Project $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in H} \|\boldsymbol{w} - \boldsymbol{y}_{t+1}\|$

Return $\bar{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$

Analysis based on OGD, with $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$.

- SGD tries to minimize $f(\boldsymbol{w}) = L_{\mathcal{D}}(\boldsymbol{w})$
- Note that $\nabla f(\boldsymbol{w}_t) = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\nabla \ell(\boldsymbol{w}_t, z)] = \mathbb{E}\mathrm{xp}_{z \sim \mathcal{D}}[\boldsymbol{v}_t] = \nabla f_t(\boldsymbol{w}_t)$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(w) = v_t w$

- SGD tries to minimize $f(w) = L_{\mathcal{D}}(w)$
- Note that $\nabla f(w_t) = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\nabla \ell(w_t, z)] = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[v_t] = \nabla f_t(w_t)$

$$\mathbb{E}\text{xp}_{z_1, \dots, z_T}[f(\bar{w})] - f(w^*) \le \mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \frac{f(w_t) - f(w^*)}{T}\right]$$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$

- SGD tries to minimize $f(\boldsymbol{w}) = L_{\mathcal{D}}(\boldsymbol{w})$
- Note that $\nabla f(\boldsymbol{w}_t) = \mathbb{E}\mathrm{xp}_{z\sim\mathcal{D}}[\nabla \ell(\boldsymbol{w}_t, z)] = \mathbb{E}\mathrm{xp}_{z\sim\mathcal{D}}[\boldsymbol{v}_t] = \nabla f_t(\boldsymbol{w}_t)$

$$\begin{aligned}\mathbb{E}\mathrm{xp}_{z_1,\ldots,z_T}[f(\bar{\boldsymbol{w}})] - f(\boldsymbol{w}^*) &\leq \mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} \frac{f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)}{T}\right] \\ &\leq \frac{1}{T}\mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} \nabla f(\boldsymbol{w}_t)(\boldsymbol{w}_t - \boldsymbol{w}^*)\right]\end{aligned}$$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$

- SGD tries to minimize $f(\boldsymbol{w}) = L_{\mathcal{D}}(\boldsymbol{w})$
- Note that $\nabla f(\boldsymbol{w}_t) = \mathbb{E}\text{xp}_{z\sim\mathcal{D}}[\nabla \ell(\boldsymbol{w}_t, z)] = \mathbb{E}\text{xp}_{z\sim\mathcal{D}}[\boldsymbol{v}_t] = \nabla f_t(\boldsymbol{w}_t)$

$$
\begin{aligned}
\mathbb{E}\text{xp}_{z_1,\ldots,z_T}[f(\bar{\boldsymbol{w}})] - f(\boldsymbol{w}^*) &\le \mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \frac{f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)}{T}\right] \\
&\le \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \nabla f(\boldsymbol{w}_t)(\boldsymbol{w}_t - \boldsymbol{w}^*)\right] \\
&= \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*)\right]
\end{aligned}
$$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(w) = v_t w$

- SGD tries to minimize $f(w) = L_{\mathcal{D}}(w)$
- Note that $\nabla f(w_t) = \mathbb{E}\mathrm{xp}_{z\sim\mathcal{D}}[\nabla \ell(w_t, z)] = \mathbb{E}\mathrm{xp}_{z\sim\mathcal{D}}[v_t] = \nabla f_t(w_t)$

$$
\begin{aligned}
\mathbb{E}\mathrm{xp}_{z_1,\ldots,z_T}[f(\bar{w})] - f(w^*) &\leq \mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} \frac{f(w_t) - f(w^*)}{T}\right] \\
&\leq \frac{1}{T}\mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} \nabla f(w_t)(w_t - w^*)\right] \\
&= \frac{1}{T}\mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} v_t(w_t - w^*)\right] \\
&= \frac{1}{T}\mathbb{E}\mathrm{xp}\left[\sum_{t=1}^{T} f_t(w_t) - f_t(w^*)\right]
\end{aligned}
$$

Stochastic Gradient Descent: Analysis

Analysis based on OGD, with $f_t(\boldsymbol{w}) = \boldsymbol{v}_t \boldsymbol{w}$

- SGD tries to minimize $f(\boldsymbol{w}) = L_{\mathcal{D}}(\boldsymbol{w})$
- Note that $\nabla f(\boldsymbol{w}_t) = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\nabla \ell(\boldsymbol{w}_t, z)] = \mathbb{E}\text{xp}_{z \sim \mathcal{D}}[\boldsymbol{v}_t] = \nabla f_t(\boldsymbol{w}_t)$

$$
\begin{aligned}
\mathbb{E}\text{xp}_{z_1,\ldots,z_T}[f(\bar{\boldsymbol{w}})] - f(\boldsymbol{w}^*) &\le \mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \frac{f(\boldsymbol{w}_t) - f(\boldsymbol{w}^*)}{T}\right] \\
&\le \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \nabla f(\boldsymbol{w}_t)(\boldsymbol{w}_t - \boldsymbol{w}^*)\right] \\
&= \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} \boldsymbol{v}_t(\boldsymbol{w}_t - \boldsymbol{w}^*)\right] \\
&= \frac{1}{T}\mathbb{E}\text{xp}\left[\sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)\right] \\
&\le \frac{\text{Regret}_{OGD}(T)}{T} \le \frac{BG}{\sqrt{T}}
\end{aligned}
$$