# CMPT 409/981: Optimization for Machine Learning

Lecture 14

Sharan Vaswani

October 29, 2024

# Recap

- For $G$-Lipschitz functions, for all $x, y \in \mathcal{D}$, $|f(y) - f(x)| \leq G \|x - y\|$. Equivalently, $\|\nabla f(w)\| \leq G$. *Example*: Hinge loss: $f(w) = \max\{0, 1 - y\langle w, x\rangle\}$ is $\|y\, x\|$-Lipschitz.
- **Subgradient**: For a convex function $f$, the subgradient of $f$ at $x \in \mathcal{D}$ is a vector $g$ that satisfies the inequality for all $y$, $f(y) \geq f(x) + \langle g, y - x\rangle$. *Example*: For $f(w) = |w|$ at $w = 0$, vectors with slope in $[-1, 1]$ and passing through the origin are subgradients.
- **Subdifferential**: The set of subgradients of $f$ at $w \in \mathcal{D}$ is referred to as the subdifferential and denoted by $\partial f(w)$. Formally, $\partial f(w) = \{g | \forall y \in \mathcal{D}; f(y) \geq f(w) + \langle g, y - w\rangle\}$.
- For unconstrained minimization of convex, non-smooth functions, $w^*$ is the minimizer of $f$ iff $0 \in \partial f(w^*)$ (this is analogous to the smooth case).
- For Lipschitz functions, we cannot relate the subgradient norm to the suboptimality in the function values. *Example*: For $f(w) = |w|$, for all $w > 0$ (including $w = 0^+$), $\|g\| = 1$.
- **Projected Subgradient Descent**: $w_{k+1} = \Pi_{\mathcal{D}}[w_k - \eta_k g_k]$, where $g_k \in \partial f(w_k)$.
- Since the sub-gradient norm does not necessarily decrease closer to the solution, to converge to the minimizer, we need to explicitly decrease the step-size.

# Minimizing convex, Lipschitz functions using Subgradient Descent

For simplicity, let us assume that $\mathcal{D} = \mathbb{R}^d$ and analyze the convergence of subgradient descent.

**Claim**: For $G$-Lipschitz, convex functions, for $\eta > 0$, $T$ iterations of subgradient descent with $\eta_k = \eta/\sqrt{k}$ converges as follows, where $\bar{w}_T = \sum_{k=0}^{T-1} w_k/T$,

$$f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[ \frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \left[1 + \log(T)\right]}{2} \right].$$

**Proof**: Similar to the previous proofs, using the update $w_{k+1} = w_k - \eta_k g_k$ where $g_k \in \partial f(w_k)$,

$$\begin{aligned}
\|w_{k+1} - w^*\|^2 &= \|w_k - w^*\|^2 - 2\eta_k \langle g_k, w_k - w^* \rangle + \eta_k^2 \|g_k\|^2 \\
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + \eta_k^2 \|g_k\|^2 \\
&\qquad \text{(Definition of subgradient with } x = w_k,\ y = w^*\text{)} \\
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + \eta_k^2\, G^2 \\
&\qquad \text{(Since } f \text{ is } G\text{-Lipschitz)}
\end{aligned}$$

$$\implies \eta_k [f(w_k) - f(w^*)] \leq \frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} + \frac{\eta_k^2\, G^2}{2}$$

# Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that $\eta_k[f(w_k) - f(w^*)] \leq \frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} + \frac{\eta_k^2 G^2}{2}$,

$$\implies \eta_{\min} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \sum_{k=0}^{T-1} \left[ \frac{\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2}{2} \right] + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \frac{\eta}{\sqrt{T}} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2 \eta^2}{2} \sum_{k=0}^{T-1} \frac{1}{k} \qquad (\text{Since } \eta_k = \eta/\sqrt{k})$$

$$\implies \frac{\sum_{k=0}^{T-1} [f(w_k) - f(w^*)]}{T} \leq \frac{1}{\sqrt{T}} \left[ \frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2} \right]$$

$$\implies f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[ \frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2} \right]$$

(Using Jensen's inequality on the LHS, and by definition of $\bar{w}_T$.)

# Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that $f(\bar{w}_T) - f(w^*) \leq \frac{1}{\sqrt{T}} \left[\frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 \eta \, [1 + \log(T)]}{2}\right]$. The above proof works for any value of $\eta$ and we can modify the proof to set the "best" value of $\eta$.

For this, let us use a constant step-size $\eta_k = \eta$. Following the same proof as before,

$$\eta_{\min} \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2} + \frac{G^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \sum_{k=0}^{T-1} [f(w_k) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{2\eta} + \frac{G^2 T \eta}{2} \qquad \text{(Since } \eta_k = \eta\text{)}$$

Setting $\eta = \frac{\|w_0 - w^*\|}{G\sqrt{T}}$, dividing by $T$ and using Jensen's inequality on the LHS,

$$f(\bar{w}_T) - f(w^*) \leq \frac{G \, \|w_0 - w^*\|}{\sqrt{T}}$$

For Lipschitz, convex functions, the above $O(1/\epsilon^2)$ rate is optimal, but we require knowledge of $G, \|w_0 - w^*\|, T$ to set the step-size.

# Minimizing convex, Lipschitz functions using Subgradient Descent

Recall that for smooth, convex functions, we could use Nesterov acceleration to obtain a faster $O(1/\sqrt{\epsilon})$ rate. On the other hand, for Lipschitz, convex functions, subgradient descent is optimal.

In order to get the $\frac{G\|w_0 - w^*\|}{\sqrt{T}}$ rate, we needed knowledge of $G$ and $\|w_0 - w^*\|$ to set the step-size. There are various techniques to set the step-size in an adaptive manner.

- AdaGrad [DHS11] is adaptive to $G$, but still requires knowing a quantity related $\|w_0 - w^*\|$ to select the "best" step-size. This influences the practical performance of AdaGrad.
- Polyak step-size [HK19] attains the desired rate without knowledge of $G$ or $\|w_0 - w^*\|$, but requires knowing $f^*$.
- Coin-Betting [OP16] does not require knowledge of $\|w_0 - w^*\|$. It only requires an estimate of $G$ and is robust to its misspecification in theory (but not quite in practice).

# Minimizing convex, Lipschitz functions using Subgradient Descent

For Lipschitz, strongly-convex functions, subgradient descent attains an $\Theta\left(\frac{1}{\epsilon}\right)$ rate. For this, the step-size depends on $\mu$ and the proof is similar to the one in (Slide 6, Lecture 10).

Subgradient descent is also optimal for Lipschitz, strongly-convex functions.

For Lipschitz functions, the convergence rates for SGD are the same as GD (with similar proofs).

| Function class | $L$-smooth + convex | $L$-smooth + $\mu$-strongly convex | $G$-Lipschitz + convex | $G$-Lipschitz + $\mu$-strongly convex |
|---|---|---|---|---|
| GD | $O\left(1/\epsilon\right)$ | $O\left(\kappa \log\left(1/\epsilon\right)\right)$ | $\Theta\left(1/\epsilon^2\right)$ | $\Theta\left(1/\epsilon\right)$ |
| SGD | $\Theta\left(1/\epsilon^2\right)$ | $\Theta\left(1/\epsilon\right)$ | $\Theta\left(1/\epsilon^2\right)$ | $\Theta\left(1/\epsilon\right)$ |

**Table 1:** Number of iterations required for obtaining an $\epsilon$-sub-optimality.

# Questions?

**Online Optimization**

Online Optimization ($w_0$, Algorithm $\mathcal{A}$, Convex set $\mathcal{C}$)
**for** $k = 1, \dots, T$ **do**
    Algorithm $\mathcal{A}$ chooses point (decision) $w_k \in \mathcal{C}$
    Environment chooses and reveals the (potentially adversarial) loss function $f_k : \mathcal{C} \to \mathbb{R}$
    Algorithm suffers a cost $f_k(w_k)$
**end for**

*Application*: **Prediction from Expert Advice**: Given $n$ experts, $\mathcal{C} = \Delta_n = \{w_i | w_i \geq 0 \, ; \, \sum_{i=1}^{n} w_i = 1\}$ and $f_k(w_k) = \langle c_k, w_k \rangle$ where $c_k \in \mathbb{R}^n$ is the loss vector.

*Application*: **Imitation Learning**: Given access to an expert that knows what action $a \in [A]$ to take in each state $s \in [S]$, learn a policy $\pi : [S] \to [A]$ that imitates the expert, i.e. we want that $\pi(a|s) \approx \pi_{\text{expert}}(a|s)$. Here, $w = \pi$ and $\mathcal{C} = \Delta_A \times \Delta_A \dots \Delta_A$ (simplex for each state) and $f_k$ is a measure of discrepancy between $\pi_k$ and $\pi_{\text{expert}}$.

# Online Optimization

- Recall that the sequence of losses $\{f_k\}_{k=1}^T$ is potentially adversarial and can also depend on $w_k$.
- **Objective**: Do well against the *best fixed decision in hindsight*, i.e. if we knew the entire sequence of losses beforehand, we would choose $w^* := \arg\min_{w \in \mathcal{C}} \sum_{k=1}^T f_k(w)$.
- **Regret**: For any fixed decision $u \in \mathcal{C}$,

$$R_T(u) := \sum_{k=1}^{T} [f_k(w_k) - f_k(u)]$$

When comparing against the best decision in hindsight,

$$R_T := \sum_{k=1}^{T} [f_k(w_k)] - \min_{w \in \mathcal{C}} \sum_{k=1}^{T} f_k(w).$$

- We want to design algorithms that achieve a *sublinear regret* (that grows as $o(T)$). A sublinear regret implies that the performance of our sequence of decisions is approaching that of $w^*$.

# Online Convex Optimization

- **Online Convex Optimization** (OCO): When the losses $f_k$ are (strongly) convex loss functions.

*Example 1*: In prediction with expert advice, $f_k(w) = \langle c_k, w \rangle$ is a linear function.

*Example 2*: In imitation learning, $f_k(\pi) = \mathbb{E}_{s \sim d^{\pi_k}}[\mathrm{KL}(\pi(\cdot|s) \,||\, \pi_{\text{expert}}(\cdot|s)]$ where $d^{\pi_k}$ is a distribution over the states induced by running policy $\pi_k$.

*Example 3*: In online control such as LQR (linear quadratic regulator) with unknown costs/perturbations, $f_k$ is quadratic.

- In Examples 2-3, the loss at iteration $k+1$ depends on the *learner*'s decision at iteration $k$.

# Online Convex Optimization

- **Online-to-Batch conversion**: If the sequence of loss functions is i.i.d from some fixed distribution, we can convert the regret guarantees into the traditional convergence guarantees for the resulting algorithm.

Formally, if $f_k$ are convex and $R(T) = O(\sqrt{T})$, then taking the expectation w.r.t the distribution generating the losses,

$$\mathbb{E}\left[\frac{R_T}{T}\right] = \mathbb{E}\left[\frac{\sum_{k=1}^{T}[f_k(w_k)] - \sum_{k=1}^{T} f_k(w^*)}{T}\right] \geq \sum_{k=1}^{T}[f(\bar{w}_T) - f(w^*)] = O\left(\frac{1}{\sqrt{T}}\right)$$

where $f(w) := \mathbb{E}[f_k(w)]$ (since the losses are i.i.d) and $\bar{w}_T := \frac{\sum_{k=1}^{T} w_k}{T}$ (since the losses are convex, we used Jensen's inequality).

- If the distribution generating the losses is a uniform discrete distribution on $n$ fixed data-points, then $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ and we are back in the finite-sum minimization setting.

- Hence, algorithms that attain $R(T) = O(\sqrt{T})$ can result in an $O\left(\frac{1}{\sqrt{T}}\right)$ convergence (in terms of the function values) for convex losses.

**Questions?**

# Online Gradient Descent

The simplest algorithm that results in sublinear regret for OCO is *Online Gradient Descent*.

**Online Gradient Descent** (OGD): At iteration $k$, the algorithm chooses the point $w_k$. After the loss function $f_k$ is revealed, OGD suffers a cost $f_k(w_k)$ and uses the function to compute

$$w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)]$$

where $\Pi_C[x] = \arg\min_{y \in \mathcal{C}} \frac{1}{2} \|y - x\|^2$.

**Claim**: If the convex set $\mathcal{C}$ has a diameter $D$ i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence losses such that each $f_k$ is convex and differentiable, OGD with a non-increasing sequence of step-sizes i.e. $\eta_k \leq \eta_{k-1}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

# Online Gradient Descent - Convex functions

**Proof**: Using the update $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)]$. Since $u \in \mathcal{C}$,

$$\|w_{k+1} - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - \Pi_{\mathcal{C}}[u]\|^2$$

Since projections are non-expansive i.e. for all $x, y$, $\|\Pi_{\mathcal{C}}[y] - \Pi_{\mathcal{C}}[x]\| \leq \|y - x\|$,

$$\begin{aligned}
&\leq \|w_k - \eta_k \nabla f_k(w_k) - u\|^2 \\
&= \|w_k - u\|^2 - 2\eta_k \langle \nabla f_k(w_k), w_k - u\rangle + \eta_k^2 \|\nabla f_k(w_k)\|^2 \\
&\leq \|w_k - u\|^2 - 2\eta_k [f_k(w_k) - f_k(u)] + \eta_k^2 \|\nabla f_k(w_k)\|^2
\end{aligned}$$

(Since $f_k$ is convex)

$$\implies 2\eta_k [f_k(w_k) - f_k(u)] \leq [\|w_k - u\|^2 - \|w_{k+1} - u\|^2] + \eta_k^2 \|\nabla f_k(w_k)\|^2$$

$$\implies R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k}\right] + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

# Online Gradient Descent - Convex functions

Recall that $R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k}\right] + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$.

$$\sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k}\right]$$

$$= \sum_{k=2}^{T} \left[\|w_k - u\|^2 \underbrace{\left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right)}_{\text{Non-negative since } \eta_k \leq \eta_{k-1}}\right] + \frac{\|w_1 - u\|^2}{2\eta_1} - \frac{\|w_{T+1} - u\|^2}{2\eta_T}$$

$$\leq D^2 \sum_{k=2}^{T} \left[\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right] + \frac{D^2}{2\eta_1} = D^2 \left[\frac{1}{2\eta_T} - \frac{1}{2\eta_1}\right] + \frac{D^2}{2\eta_1} = \frac{D^2}{2\eta_T}$$

(Since $\|x - y\| \leq D$ for all $x, y \in \mathcal{C}$)

Putting everything together,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

# Online Gradient Descent - Convex, Lipschitz functions

**Claim**: If the convex set $\mathcal{C}$ has a diameter $D$ i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence losses such that each $f_k$ is convex, differentiable and $G$-Lipschitz, OGD with $\eta_k = \frac{\eta}{\sqrt{k}}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta$$

**Proof**: Since the step-size is decreasing, we can use the general result from the previous slide,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \left\|\nabla f_k(w_k)\right\|^2 \leq \frac{D^2}{2\eta_T} + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k \qquad \text{(Since } f_k \text{ is } G\text{-Lipschitz)}$$

$$\implies R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + \frac{G^2 \eta}{2} \sum_{k=1}^{T} \frac{1}{\sqrt{k}} \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta \qquad \text{(Since } \textstyle\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \leq 2\sqrt{T}\text{)}$$

In order to find the "best" $\eta$, set it such that $D^2/2\eta = G^2 \eta$, implying that $\eta = D/\sqrt{2}G$ and $R_T(u) \leq \sqrt{2}\, DG \sqrt{T}$. Hence, OGD with a decreasing step-size attains sublinear $\Theta(\sqrt{T})$ regret for convex, Lipschitz functions.

# Online Gradient Descent - Strongly-convex, Lipschitz functions

**Claim**: If the convex set $\mathcal{C}$ has a diameter $D$, for an arbitrary sequence losses such that each $f_k$ is $\mu_k$ strongly-convex (s.t. $\mu := \min_{k \in [T]} \mu_k > 0$), $G$-Lipschitz and differentiable, then OGD with $\eta_k = \frac{1}{\sum_{i=1}^{k} \mu_i}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{G^2}{2\mu} \left(1 + \log(T)\right)$$

**Proof**: Similar to the convex proof, use the update $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)]$. Since $u \in \mathcal{C}$,

$$\begin{aligned}
\|w_{k+1} - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - u\|^2 &= \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - \Pi_{\mathcal{C}}[u]\|^2 \\
&\leq \|w_k - u\|^2 - 2\eta_k \langle \nabla f_k(w_k), w_k - u \rangle + \eta_k^2 \|\nabla f_k(w_k)\|^2 \\
&\leq \|w_k - u\|^2 (1 - \mu_k \eta_k) - 2\eta_k [f_k(w_k) - f_k(u)] + \eta_k^2 \|\nabla f_k(w_k)\|^2
\end{aligned}$$

(Since $f_k$ is $\mu_k$ strongly-convex)

$$\implies R_T(u) \leq \sum_{k=1}^{T} \left[ \frac{\|w_k - u\|^2 (1 - \mu_k \eta_k) - \|w_{k+1} - u\|^2}{2\eta_k} \right] + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k$$

(Since $f_k$ is $G$-Lipschitz)

# Online Gradient Descent - Strongly-convex, Lipschitz functions

Recall that $R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 (1-\mu_k \eta_k) - \|w_{k+1} - u\|^2}{2\eta_k}\right] + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k$.

$$\sum_{k=1}^{T} \left[\frac{\left\|w_k - u\right\|^2 (1 - \mu_k \eta_k) - \left\|w_{k+1} - u\right\|^2}{2\eta_k}\right]$$

$$= \sum_{k=2}^{T} \left[\left\|w_k - u\right\|^2 \underbrace{\left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}} - \frac{\mu_k}{2}\right)}_{=0}\right] + \left\|w_1 - u\right\|^2 \underbrace{\left[\frac{1}{2\eta_1} - \frac{\mu_1}{2}\right]}_{=0} - \frac{\left\|w_{T+1} - u\right\|^2}{2\eta_T} \leq 0$$

(Since $\eta_k = \frac{1}{\sum_{i=1}^{k} \mu_i}$)

Putting everything together,

$$R_T(u) \leq \frac{G^2}{2} \sum_{k=1}^{T} \frac{1}{\mu k} \leq \frac{G^2}{2\mu} (1 + \log(T))$$

(Since $\mu := \min_{k \in [T]} \mu_k$ and $\sum_{k=1}^{T} 1/k \leq 1 + \log(T)$)

**Lower Bound**: There is an $\Omega(\log(T))$ lower-bound on the regret for strongly-convex, Lipschitz functions and hence OGD is optimal in this setting!

**Questions?**

# References i

John Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization.*, Journal of machine learning research **12** (2011), no. 7.

Elad Hazan and Sham Kakade, *Revisiting the polyak step size*, arXiv preprint arXiv:1905.00313 (2019).

Francesco Orabona and Dávid Pál, *Coin betting and parameter-free online learning*, Advances in Neural Information Processing Systems **29** (2016).