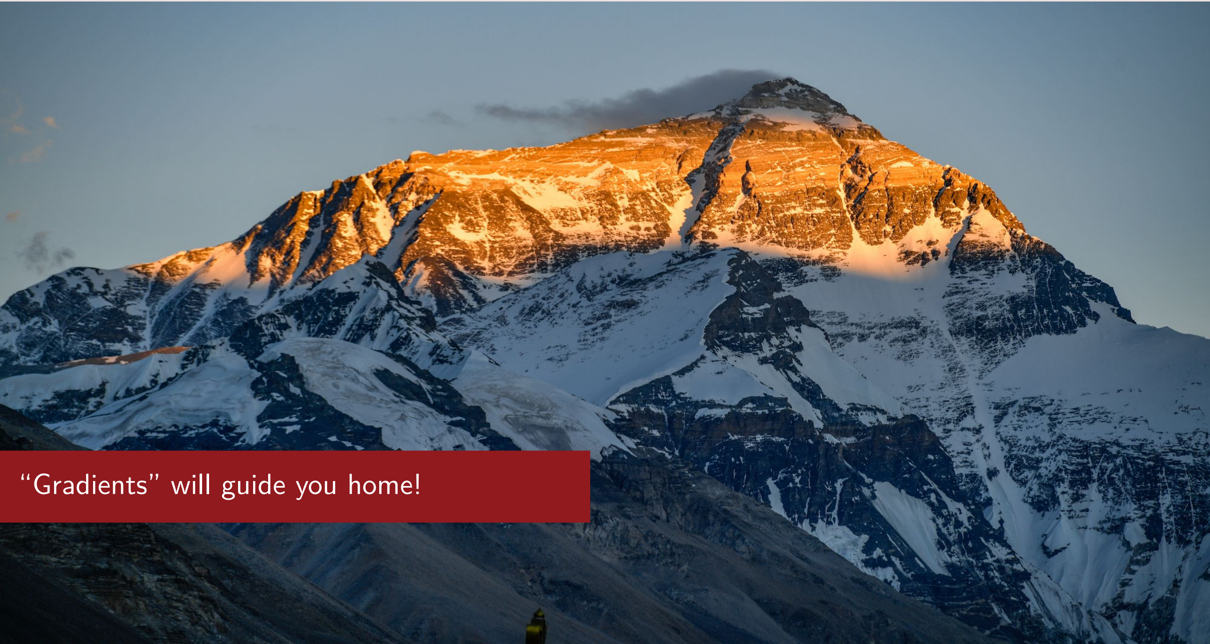
Optimization Algorithms in Machine Learning

Nirupam Gupta

Department of Computer Science

UNIVERSITY OF COPENHAGEN



“Gradients” will guide you home!

Logistic regression

Consider a dataset $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{0, 1\}$ for all i.

For $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$, **define** $z_i = w^\intercal x_i + b$ and $p_i = \text{Sigmoid}(z_i) = \frac{1}{1+\exp(-z_i)}$.

p_i is the probability that the true label is 1 for x_i for a model parameterized by $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$.

Cross-entropy loss. For each sample (x_i, y_i), define

$$\ell_i(w, b) = -\left[y_i \log\left(\frac{1}{1+\exp(-z_i)}\right) + (1 - y_i)\log\left(\frac{\exp(-z_i)}{1+\exp(-z_i)}\right)\right]$$

Optimization problem.

$$\underset{w \in \mathbb{R}^m, b \in \mathbb{R}}{\text{Minimize}} \quad \mathcal{L}(w, b) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(w, b)$$

The loss function $\mathcal{L}(w, b)$ is differentiable and convex (why?).

Unconstrained minimization

Optimization problem:

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

If f is differentiable then by Taylor's **first-order approximation**:

$$f(w') = f(w) + \langle \nabla f(w), w' - w \rangle + o\left(\|w' - w\|\right),$$

where $\lim_{r \to 0} \frac{o(r)}{r} = 0$ and $o(0) = 0$.

Descent methods. Initial guess w_1, updated iteratively as $w_{t+1} = w_t + u_t$ to generate a *relaxation sequence* $\{f(w_t)\}_{t=1}^{\infty}$, i.e., $f(w_{t+1}) \leq f(w_t)$.

If $f(w)$ is lower bounded for all $w \in \mathbb{R}^d$, the above sequence converges.

Gradient descent

Method of gradient descent. For $\gamma_t \in \mathbb{R}_{++}$, update rule: $w_{t+1} = w_t - \gamma_t \nabla f(w_t)$.

$$f(w_{t+1}) = f(w_t) - \gamma_t \left\| \nabla f(w_t) \right\|^2 + o\left(\gamma_t \left\| \nabla f(w_t) \right\|\right).$$

For small enough γ_t, we have: (why?)

$$f(w_{t+1}) \leq f(w_t) - c_t \, \gamma_t \left\| \nabla f(w_t) \right\|^2 \leq f(w_t),$$

for some $c_t \in (0, 1]$.

$f(w_{t+1}) = f(w_t) \iff \nabla f(w_t) = 0$, and w_t is referred to as a **stationary point**.

If f is a convex function, gradient descent method converges to a minimum point.

Rate of convergence: Lipschitzness

Approximate stationarity. How many iterations until $\|\nabla f(w_t)\| \leq \varepsilon$?

This generally depends on how "nicely" can the *residue* $o(\cdot)$ be bounded.

Lipschitz smoothness. There exists $L \in \mathbb{R}_+$ such that $\|\nabla f(w) - \nabla f(w')\| \leq L \|w - w'\|$. In that case,

$$o(\|w' - w\|) \leq \frac{L}{2} \|w' - w\|^2 .$$

Thus, method of gradient descent yields,

$$f(w_{t+1}) \leq f(w_t) - \gamma_t \left(1 - \gamma_t \frac{L}{2}\right) \|\nabla f(w_t)\|^2 .$$

Rate of convergence: choosing the right step-size

Method of gradient descent. For $\gamma_t \in \mathbb{R}_{++}$, update rule: $w_{t+1} = w_t - \gamma_t \nabla f(w_t)$.

Under L-Lipschitz smoothness, we obtain that

$$f(w_{t+1}) \leq f(w_t) - \gamma_t \left(1 - \gamma_t \frac{L}{2}\right) \left\|\nabla f(w_t)\right\|^2 .$$

- **Constant step-size.** For all iterations t, $\gamma_t = \gamma$
- **Diminishing step-size.** For all iterations t, $\gamma_t = \frac{\gamma_o}{\sqrt{t}}$
- **Armijo rule.** For each iteration t, determine γ_t such that

$$f(w_t) - \beta \gamma_t \left\|\nabla f(w_t)\right\|^2 \leq f(w_{t+1}) \leq f(w_t) - \alpha \gamma_t \left\|\nabla f(w_t)\right\|^2 ,$$

where $0 < \alpha < \beta < 1$.

Rate of convergence: constant learning rate

$$f(w_{t+1}) \leq f(w_t) - \gamma \left(1 - \gamma \tfrac{L}{2}\right) \left\|\nabla f(w_t)\right\|^2 .$$

Suppose $\gamma \leq \frac{1}{L}$. Then,

$$f(w_{t+1}) \leq f(w_t) - \tfrac{\gamma}{2} \left\|\nabla f(w_t)\right\|^2 .$$

Therefore,

$$\tfrac{\gamma}{2} \textstyle\sum_{t=1}^{T} \left\|\nabla f(w_t)\right\|^2 \leq f(w_1) - f(w_{T+1}) \leq f(w_1) - f^*,$$

where $f^* = \min f(w)$. Hence, $\frac{1}{T} \sum_{t=1}^{T} \left\|\nabla f(w_t)\right\|^2 \leq \frac{2}{T}(f(w_1) - f^*)$, which implies that

$$\min_{t \in [T]} \left\|\nabla f(w_t)\right\| \leq \sqrt{\tfrac{2}{T}(f(w_1) - f^*)} \in \mathcal{O}\left(\tfrac{1}{\sqrt{T}}\right) .$$

Gradient descent under Lipschitz smoothness & convexity

When f is convex (and L-Lipszhitz smooth), $\langle \nabla f(w) - \nabla f(w'), \, w - w' \rangle \geq \frac{1}{L} \left\| \nabla f(w) - \nabla f(w') \right\|^2$.

Let w^* be a minimizer of $f(w)$. Then,

$$\begin{aligned} \left\| w_{t+1} - w^* \right\|^2 &\leq \left\| w_t - w^* \right\|^2 - 2\gamma \left\langle \nabla f(w_t), \, w_t - w^* \right\rangle + \gamma^2 \left\| \nabla f(w_t) \right\|^2 \\ &= \left\| w_t - w^* \right\|^2 - \gamma \left(\frac{2}{L} - \gamma \right) \left\| \nabla f(w_t) \right\|^2 . \end{aligned}$$

Suppose $\gamma \leq \frac{1}{L}$. Then, $\left\| w_t - w^* \right\|^2 \leq \left\| w_1 - w^* \right\|^2$. Recall that, under L-Lipschitz smoothness,

$$f(w_{t+1}) \leq f(w_t) - \frac{\gamma}{2} \left\| \nabla f(w_t) \right\|^2 .$$

Due to convexity, $f(w_t) - f^* \leq \langle \nabla f(w_t), \, w_t - w^* \rangle \leq \left\| w_t - w^* \right\| \left\| \nabla f(w_t) \right\| \leq \left\| w_1 - w^* \right\| \left\| \nabla f(w_t) \right\|$. Thus,

$$f(w_{t+1}) \leq f(w_t) - \frac{\gamma}{2 \left\| w_1 - w^* \right\|^2} \left(f(w_t) - f^* \right)^2 .$$

From above we obtain that (why?), $f(w_T) - f^* \in \mathcal{O}\left(\frac{1}{T}\right)$.

Gradient descent under Lipschitz smoothness & strong convexity

When f is μ-strongly convex, $2\mu(f(w) - f^*) \leq \|\nabla f(w)\|^2$.

Recall that, under L-Lipschitz smoothness, when $\gamma \leq \frac{1}{L}$ we have:

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

Therefore, under strong convexity,

$$f(w_{t+1}) - f^* \leq (1 - \mu\gamma) (f(w_t) - f^*) .$$

Hence (why?), $f(w_T) - f^* \in \mathcal{O}\left(\exp\left(-\frac{1}{\kappa} T\right)\right)$, where $\kappa = \frac{L}{\mu}$ is the *condition number* of $f(w)$.

Method of stochastic gradient descent (SGD)

In ML the loss function is the sum of point-wise loss functions: $\mathcal{L}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(w)$.

Gradient descent does not scale well with n. A more practical approach:

$$g_t = \frac{1}{b}\sum_{i\in B} \nabla \ell_i(w_t),$$

where B is a random subset of S called a **batch** of size b called the **batch-size**.

Given w_t, we have: $\mathbb{E}[g_t] = \nabla \mathcal{L}(w_t)$ and we assume: $\mathbb{E}\left[\|\nabla \ell_i(w_t) - \mathcal{L}(w_t)\|^2\right] \le \sigma^2$.

SGD update rule: $w_{t+1} = w_t - \gamma_t g_t$.

SDG under Lipschitz smoothness & strong convexity

Due to L-Lipchitz smoothness, $\mathcal{L}(w_{t+1}) \leq \mathcal{L}(w_t) - \gamma_t \langle \nabla \mathcal{L}(w_t), g_t \rangle + \gamma_t^2 \frac{L}{2} \|g_t\|^2$.

Given $w_1, \ldots, w_t$, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(w_{t+1})\right] &\leq \mathcal{L}(w_t) - \gamma_t \langle \nabla \mathcal{L}(w_t), \mathbb{E}\left[g_t\right] \rangle + \gamma_t^2 \frac{L}{2} \mathbb{E}\left[\|g_t\|^2\right] \\
&\leq \mathcal{L}(w_t) - \gamma_t \|\nabla \mathcal{L}(w_t)\|^2 + \gamma_t^2 \tfrac{L}{2} \left(\tfrac{\sigma^2}{b} + \|\nabla \mathcal{L}(w_t)\|^2 \right) .
\end{aligned}$$

Under μ-strong convexity, $2\mu(\mathcal{L}(w) - \mathcal{L}^*) \leq \|\nabla \mathcal{L}(w)\|^2$. Thus, if $\gamma_t \leq \frac{1}{L}$, then

$$\mathbb{E}\left[\mathcal{L}(w_{t+1}) - \mathcal{L}^*\right] \leq (1 - \mu\gamma_t)\, \mathbb{E}\left[\mathcal{L}(w_t) - \mathcal{L}^*\right] + \tfrac{L\sigma^2}{2b}\gamma_t^2 .$$

If $\gamma_t = \gamma$, this implies, (why?) $\mathbb{E}\left[\mathcal{L}(w_{T+1}) - \mathcal{L}^*\right] \leq (1 - \mu\gamma)^T\, \mathbb{E}\left[\mathcal{L}(w_1) - \mathcal{L}^*\right] + \frac{L\sigma^2}{2\mu b}\gamma$.

Substituting $\gamma = \frac{\log T}{T}$, we obtain that $\mathbb{E}\left[\mathcal{L}(w_{T+1}) - \mathcal{L}^*\right] \in \widetilde{\mathcal{O}}\left(\kappa \frac{\sigma^2}{b} \cdot \frac{1}{T}\right)$.

Logistic regression with l_2-regulizer

Consider a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{0, 1\}$ for all i.

For $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$, **define** $z_i = w^\intercal x_i + b$ and $p_i = \text{Sigmoid}(z_i) = \frac{1}{1+\exp(-z_i)}$.

Cross-entropy loss. For each sample (x_i, y_i), define

$$\ell_i(w, b) = -\left[y_i \log\left(\frac{1}{1+\exp(-z_i)}\right) + (1 - y_i)\log\left(\frac{\exp(-z_i)}{1+\exp(-z_i)}\right)\right]$$

Regularized ERM:.

$$\underset{w\in\mathbb{R}^m, b\in\mathbb{R}}{\text{Minimize}} \quad \mathcal{L}(w, b) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(w, b) + \frac{\mu}{2}(\|w\|^2 + b^2)$$

The loss function $\mathcal{L}(w, b)$ is μ-strongly convex (why?).

References & further readings

The lecture notes are based on Sections 2 - 6 of "Handbook of Convergence Theorems for (Stochastic) Gradient Methods" by Garrigos and Gower.

Additional reading:

- **Stochastic momentum:** Section 7.
- **Stochastic subgradient descent:** Section 9.