Week 2 - L04
Stochastic Gradient Descent
(Examples)

CS 295 Optimization for Machine Learning

Ioannis Panageas

Optimization in ML, SGD to the rescue

Definition (Risk Minimization). *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \text{ where } L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

Optimization in ML, SGD to the rescue

Definition (Risk Minimization). *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \text{ where } L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

Approach one:

1. Take enough (say n) samples z_i independently and consider the estimate $\bar{L}(x) \coloneqq \frac{1}{n}\sum_{\mathrm{i}} \ell(x, z_i)$. By Law of Large Numbers this is a close enough with high probability.
2. Run a first order optimization algorithm (say GD) on $\bar{L}(x)$.

Remark:
If we do not know the form of $\ell(x,z)$ and we only have oracle access it is not possible. Also many calculations per iteration…

Optimization in ML, SGD to the rescue

Definition (Risk Minimization). *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \text{ where } L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

Approach one:

Or use SGD!

2. Run a first order optimization algorithm (say GD) on $\bar{L}(x)$.

Remark:

If we do not know the form of $\ell(x,z)$ and we only have oracle access it is not possible. Also many calculations per iteration…

Optimization in ML, SGD to the rescue

Definition (Risk Minimization)**.** *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \ \textit{where}\ L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

Approach two (SGD):

1. For each iteration $t+1$, take a fresh sample z_t independently from $z_1, \ldots, z_{t-1}$ and consider the unbiased estimate $\nabla_x \ell(x_t, z_t)$.
2. Update $x_{t+1} = x_t - \alpha_t \nabla_x \ell(x_t, z_t)$.

Optimization in ML, SGD to the rescue

Definition (Risk Minimization). *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \text{ where } L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

For t:=1 to T do

1. sample $z \sim D$.
2. Pick $v_t \in \partial \ell(x_t, z)$.
3. $x_{t+1} = x_t - \alpha_t v_t$.

Return $\frac{1}{T} \sum x_t$.

Optimization in ML, SGD to the rescue

Definition (Risk Minimization). *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \text{ where } L(x) := \mathbb{E}_{z \sim D}[\ell(x,z)].$$

Remark:

What if $L(x) = \frac{1}{n}\sum_{i=1}^{n} g_i(x)$?

Less Cost per iteration (but you need to "pay" the variance):

1. For each iteration $t+1$, take uniformly at random independently index i from $1, \dots, n$ and consider the (unbiased estimate) $\nabla_x g_i(x)$.
2. Update $x_{t+1} = x_t - \alpha_t \nabla_x g_i(x)$.

An example (SGD approach)

Definition (MLE for Gaussian)**.** *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x, z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x, 1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)}[-\log p_x(z)].$$

Any guesses what is the minimizer of the above?

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x,z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x,1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)}[-\log p_x(z)].$$

Of course $x^* = \mu$. Remarks on Maximum (log)-Likelihood:

1. Standard approach for parameter estimation of parametric families of distributions, i.e., create an optimization problem!
2. Under assumptions, Maximum (log) Likelihood Estimator is consistent!
3. Above boils down to $\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z-x)^2}{2} \right].$
4. Let's do SGD…

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x, z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x, 1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z - x)^2}{2} \right].$$

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x, z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x, 1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z-x)^2}{2} \right].$$

- The derivative is just $(x - z)$ and $\mathbb{E}[(x-z)^2] = 1 + (x - \mu)^2$.
- The second derivative is 1, hence 1-strongly convex.
- Start from $x_0 = 0$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t(x_t - z_t)$.

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x,z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x,1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z-x)^2}{2} \right].$$

- The derivative is just $(x - z)$ and $\mathbb{E}[(x-z)^2] = 1 + (x-\mu)^2$.
- The second derivative is 1, hence 1-strongly convex.
- Start from $x_0 = 0$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t(x_t - z_t)$.

Choosing $a_t = \frac{1}{t}$ (check SGD thm), what is x_T?

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x, z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x, 1)$. We do not know μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z-x)^2}{2} \right].$$

- The derivative is just $(x - z)$ and $\mathbb{E}[(x-z)^2] = 1 + (x-\mu)^2$.
- The second derivative is 1, hence 1-strongly convex.
- Start from $x_0 = 0$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t(x_t - z_t)$.

Recall for $T = \Theta\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ we get error ϵ!

Choosing $a_t = \frac{1}{t}$ (check SGD thm), what is x_T?

An example (SGD approach)

Definition (MLE for Gaussian). *Let $z \sim \mathcal{N}(\mu, 1)$ and $\ell(x,z) := -\log p_x(z)$ denotes the log-likelihood of $\mathcal{N}(x,1)$. We do not know μ. We are interested in solving:*

$$\min_{x\in\mathbb{R}} \mathbb{E}_{z\sim\mathcal{N}(\mu,1)}\left[\frac{(z-x)^2}{2}\right].$$

- The derivative is just $(x-z)$ and $\mathbb{E}[(x-z)^2] = 1 + (x-\mu)^2$.
- Th

You can get ϵ-close to μ after $\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon^2}$ itearations! Not tight, why?

- Start from $x_0 = 0$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t(x_t - z_t)$.

It is the empirical mean, i.e., $x_T = \frac{1}{T}\sum_i z_i$ **!**

An example (SGD approach)

Problem (Bias of a coin). *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p-\tilde{p}| \leq \epsilon$?*

$$\boxed{\text{Hint: Density } f_p(z) = p^z(1-p)^{1-z}}$$

An example (SGD approach)

Problem (Bias of a coin). *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p - \tilde{p}| \leq \epsilon$?*

Hint: Density $f_p(z) = p^z(1-p)^{1-z}$

- A discrete probabilist will use Chernoff bounds or Chebyshev!
- A statistitian/optimization guy will solve $\min_x \mathbb{E}[-\log f_x(z)]$.

An example (SGD approach)

Problem (Bias of a coin). *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p-\tilde{p}| \leq \epsilon$?*

Hint: Density $f_p(z) = p^z(1-p)^{1-z}$

- A discrete probabilist will use Chernoff bounds or Chebyshev!
- A statistitian/optimization guy will solve $\min_x \mathbb{E}[-\log f_x(z)]$.

We would like to solve (of course $x^* = p$ is the solution but we don't know p)

$$\min_x \mathbb{E}[-z \log x - (1-z)\log(1-x)].$$

An example (SGD approach)

Problem (Bias of a coin). *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p - \tilde{p}| \leq \epsilon$?*

Hint: Density $f_p(z) = p^z(1-p)^{1-z}$

- The derivative of ℓ is just $-\frac{z}{x} + \frac{(1-z)}{1-x} = \frac{x-z}{x(1-x)}$, which is in absolute value at most $\frac{1}{\epsilon}$ for $x \in (\epsilon, 1-\epsilon)$.
- The second derivative of L is $\frac{p}{x^2} + \frac{1-p}{(1-x)^2}$, hence $4(p-p^2)$-strongly convex in $(0,1)$.
- Start from $x_0 = 1/2$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t \frac{(x_t - z_t)}{x_t(1-x_t)}$.

An example (SGD approach)

Problem (Bias of a coin). *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p - \tilde{p}| \leq \epsilon$?*

Hint: Density $f_p(z) = p^z(1-p)^{1-z}$

- The derivative of ℓ is just $-\frac{z}{x} + \frac{(1-z)}{(1-x)} = \frac{x-z}{x(1-x)}$, which is in absolute value

You can get ϵ-close to p after $\frac{1}{4(p-p^2)\epsilon^6} \ln \frac{1}{\epsilon^2}$ itearations! Not tight, why?

- The second derivative of L is $\frac{1}{x^2} + \frac{1}{(1-x)^2}$, hence $4(p-p^2)$-strongly convex in $(0,1)$.
- Start from $x_0 = 1/2$.
- At iteration $t+1$, get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t \frac{(x_t - z_t)}{x_t(1-x_t)}$.

A strange example.

Problem (Mixture of Gaussians)**.** *Assume you have access to i.i.d samples from $z \sim \mathcal{N}(\mu, 1)$. However, there is an adversary that with probability $1/2$ corrupts z and gives you $-z$. Can you infer/estimate μ?*

A strange example.

Problem (Mixture of Gaussians). *Assume you have access to i.i.d samples from $z \sim \mathcal{N}(\mu, 1)$. However, there is an adversary that with probability $1/2$ corrupts z and gives you $-z$. Can you infer/estimate μ?*

$$\text{Need to solve: } \min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[-\log \left(\frac{1}{2\sqrt{2\pi}} e^{(z-x)^2/2} + \frac{1}{2\sqrt{2\pi}} e^{(z+x)^2/2} \right) \right].$$

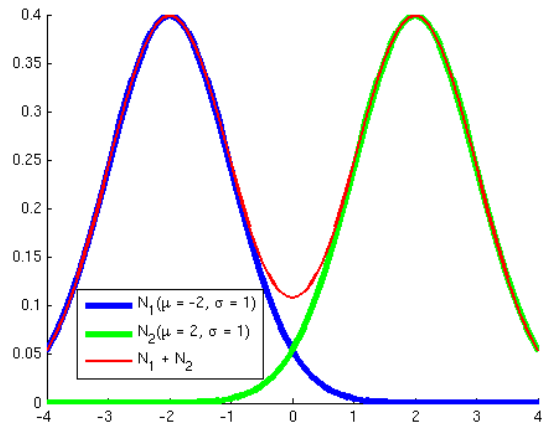
- Is it convex? **Exercise 5.**

A strange example.

Problem (Mixture of Gaussians)**.** *Assume you have access to i.i.d samples from $z \sim \mathcal{N}(\mu, 1)$. However, there is an adversary that with probability $1/2$ corrupts z and gives you $-z$. Can you infer/estimate μ?*

$$\text{Need to solve: } \min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[-\log \left(\frac{1}{2\sqrt{2\pi}} e^{(z-x)^2/2} + \frac{1}{2\sqrt{2\pi}} e^{(z+x)^2/2} \right) \right].$$

- Is it convex? **Exercise 5.**



Conclusion

- Examples on SGD:
  - MLE, testing bias of coin.
- Non-convex examples: Mixture of Gaussians

- Next week we will talk about online learning/optimization!