1 Introduction to Stochastic Gradient Descent

1.1 Problem Setup

Let $(\mathcal{X} \times \mathcal{Y}, P)$ be a probability space, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. We observe a dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ consisting of n i.i.d. samples from the distribution P.

Consider a parametric model $f : \Theta \times \mathcal{X} \to \mathcal{Y}$, where $\Theta \subseteq \mathbb{R}^d$ is the parameter space. For neural networks, $f(\theta; x)$ represents the network's output given parameters $\theta \in \Theta$ and input $x \in \mathcal{X}$.

1.2 Loss Functions

We define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ that measures the discrepancy between predictions and targets. For regression tasks, we typically use the squared loss:

$$\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|^2$$

This induces the following hierarchy of loss functions:

1. **Individual sample loss**: For a single data point (x, y),

$$\ell(\theta; x, y) := \ell(f(\theta; x), y)$$

2. **Empirical risk** (finite-sample loss): Over the dataset $\mathcal{D}_n$,

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; X_i, Y_i)$$

3. **Population risk** (expected loss): Over the true distribution P,

$$L(\theta) := \mathbb{E}_{(X,Y)\sim P}[\ell(\theta; X, Y)]$$

In practice, we only have access to $L_n(\theta)$ and use it as a proxy for $L(\theta)$, which we truly wish to minimize.

1.3 Gradients

Define the gradients with respect to parameters θ:

1. **Individual gradient**: $g_i(\theta) := \nabla_\theta \ell(\theta; X_i, Y_i)$
2. **Empirical gradient**: $g_n(\theta) := \nabla L_n(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta)$
3. **Stochastic gradient**: For a batch $\mathcal{B} \subset D_n$ of size B sampled uniformly at random,

$$g_{\mathcal{B}}(\theta) := \frac{1}{B}\sum_{i\in\mathcal{B}} g_i(\theta)$$

4. **Population gradient**: $g(\theta) := \nabla L(\theta) = \mathbb{E}_{(X,Y)\sim P}[\nabla_\theta \ell(\theta; X, Y)]$

Note that $\mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(\theta)] = \nabla L_n(\theta)$, making the stochastic gradient an unbiased estimator of the empirical gradient.

Algorithm 1 Stochastic Gradient Descent

Initialize $\theta_0 \in \Theta$
for $k = 0, 1, 2, ..., K-1$ **do**
    Sample mini-batch $\mathcal{B}_k \subseteq D_n$ uniformly at random
    Compute stochastic gradient $g_{\mathcal{B}_k}(\theta_k)$
    Update: $\theta_{k+1} = \theta_k - \eta_k g_{\mathcal{B}_k}(\theta_k)$
end for

1.4 The SGD Algorithm

Stochastic Gradient Descent (SGD) performs the following iterative update:
where $\eta_k > 0$ is the learning rate (or step size) at iteration k.

1.5 Mini-batch Sampling Strategies

There are two common sampling strategies:

1. **With replacement**: Each element of $\mathcal{B}_k$ is drawn independently and uniformly from $\{1, ..., n\}$
2. **Without replacement**: Sample B distinct indices uniformly from $\{1, ..., n\}$

The with-replacement case is often easier to analyze theoretically due to independence.

1.6 Decomposition into Drift and Noise

It is useful to decompose the SGD update as:

$$\theta_{k+1} = \theta_k - \eta_k g_n(\theta_k) + \eta_k \xi(\theta_k)$$

where the **gradient noise** is defined as:

$$\xi(\theta_k) := g_n(\theta_k) - g_{\mathcal{B}_k}(\theta_k)$$

This yields a decomposition into:

- **Drift**: $\eta_k g_n(\theta_k)$ (deterministic gradient)
- **Diffusion**: $\eta_k \xi(\theta_k)$ (stochastic perturbation)

1.7 Special Cases

1. **Full-batch gradient descent**: $\mathcal{B}_k = \{1, ..., n\}$, so $\xi_k = 0$ (deterministic)
2. **Single-sample SGD**: $|\mathcal{B}_k| = 1$, maximum noise
3. **Constant learning rate**: $\eta_k = \eta$ for all k
4. **Decaying learning rate**: $\eta_k \to 0$ as $k \to \infty$ (e.g., $\eta_k = \eta_0/\sqrt{k+1}$)

“‘latex

1.8 Properties of the Gradient Noise

The stochasticity in SGD arises from two distinct sources of randomness, which we now analyze.

1.8.1 Two Sources of Gradient Noise

1. Sampling Noise (Empirical vs. Population)

The first source of randomness comes from using a finite dataset. Even if we compute the full empirical gradient $g_n(\theta)$, it differs from the true population gradient $g(\theta)$:

$$\nu_{\text{sampling}}(\theta) := g_n(\theta) - g(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta) - \mathbb{E}_{(X,Y)\sim P}[\nabla_\theta \ell(\theta; X, Y)]$$

Since the data points are i.i.d. samples from P, we have:

$$\mathbb{E}_{\mathcal{D}_n}[\nu_{\text{sampling}}(\theta)] = \mathbb{E}_{\mathcal{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right] - g(\theta) = 0 \tag{1}$$

The covariance is:

$$\begin{aligned}
\text{Cov}_{\mathcal{D}_n}[\nu_{\text{sampling}}(\theta)] &= \text{Cov}_{\mathcal{D}_n}\left[\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)\right] && (2)\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Cov}_{(X_i,Y_i)}[g_i(\theta)] && (3)\\
&= \frac{1}{n}\Sigma_{\text{pop}}(\theta) && (4)
\end{aligned}$$

where $\Sigma_{\text{pop}}(\theta) := \text{Cov}_{(X,Y)\sim P}[\nabla_\theta \ell(\theta; X, Y)]$ is the population gradient covariance.

2. Mini-batch Noise (Empirical vs. Batch)

The second source of randomness comes from using mini-batches instead of the full dataset:

$$\nu_{\text{batch}}(\theta) := g_{\mathcal{B}}(\theta) - g_n(\theta) = \frac{1}{B}\sum_{i\in\mathcal{B}} g_i(\theta) - \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$$

For uniform sampling with replacement, conditioning on the dataset $\mathcal{D}_n$:

$$\mathbb{E}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)|\mathcal{D}_n] = \mathbb{E}_{\mathcal{B}}\left[\frac{1}{B}\sum_{i\in\mathcal{B}} g_i(\theta)\right] - g_n(\theta) = 0 \tag{5}$$

The conditional covariance is:

$$\begin{aligned}
\text{Cov}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)|\mathcal{D}_n] &= \text{Cov}_{\mathcal{B}}\left[\frac{1}{B}\sum_{i\in\mathcal{B}} g_i(\theta)\middle|\mathcal{D}_n\right] && (6)\\
&= \frac{1}{B}\Sigma_{\text{emp}}(\theta, \mathcal{D}_n) && (7)
\end{aligned}$$

where $\Sigma_{\text{emp}}(\theta, \mathcal{D}_n) := \frac{1}{n}\sum_{i=1}^{n}[g_i(\theta) - g_n(\theta)][g_i(\theta) - g_n(\theta)]^T$ is the empirical gradient covariance.

1.8.2 Total Gradient Noise

The total noise in SGD is:

$$\xi_{\text{total}}(\theta) := g_{\mathcal{B}}(\theta) - g(\theta) = \underbrace{g_{\mathcal{B}}(\theta) - g_n(\theta)}_{\nu_{\text{batch}}(\theta)} + \underbrace{g_n(\theta) - g(\theta)}_{\nu_{\text{sampling}}(\theta)}$$

However, in the SGD algorithm as defined, we use:

$$\xi(\theta) := g_n(\theta) - g_{\mathcal{B}}(\theta) = -\nu_{\text{batch}}(\theta)$$

This is because we can only compute deviations from the empirical gradient, not the unknown population gradient.

1.8.3 Covariance Structure of SGD Noise

For the noise $\xi(\theta) = g_n(\theta) - g_{\mathcal{B}}(\theta)$ used in SGD:

Proposition 1. *Under uniform mini-batch sampling with replacement, the gradient noise has the following properties:*

1. **Unbiasedness**: $\mathbb{E}_{\mathcal{B}}[\xi(\theta)|\mathcal{D}_n, \theta] = 0$
2. **Covariance**:
$$\mathrm{Cov}_{\mathcal{B}}[\xi(\theta)|\mathcal{D}_n, \theta] = \frac{1}{B}\Sigma_{\mathrm{emp}}(\theta, \mathcal{D}_n)$$
3. **Expected Covariance**: Taking expectation over datasets,
$$\mathbb{E}_{\mathcal{D}_n}[\mathrm{Cov}_{\mathcal{B}}[\xi(\theta)|\mathcal{D}_n, \theta]] = \frac{1}{B} \cdot \frac{n-1}{n}\Sigma_{\mathrm{pop}}(\theta)$$

Proof.

1. Follows immediately from $\mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(\theta)] = g_n(\theta)$.
2. For sampling with replacement:
$$\mathrm{Cov}_{\mathcal{B}}[\xi(\theta)|\mathcal{D}_n] = \mathrm{Cov}_{\mathcal{B}}[g_{\mathcal{B}}(\theta)|\mathcal{D}_n] \tag{8}$$
$$= \mathrm{Var}_{\mathcal{B}}\left[\frac{1}{B}\sum_{j=1}^{B} g_{I_j}(\theta)\middle|\mathcal{D}_n\right] \tag{9}$$
where I_j are i.i.d. uniform on $\{1, ..., n\}$. Since the I_j are independent:
$$= \frac{1}{B^2}\sum_{j=1}^{B}\mathrm{Var}_{I_j}[g_{I_j}(\theta)|\mathcal{D}_n] \tag{10}$$
$$= \frac{1}{B} \cdot \frac{1}{n}\sum_{i=1}^{n}[g_i(\theta) - g_n(\theta)][g_i(\theta) - g_n(\theta)]^T \tag{11}$$
3. Using the identity $\mathbb{E}[\Sigma_{\mathrm{emp}}] = \frac{n-1}{n}\Sigma_{\mathrm{pop}}$ completes the proof.

1.8.4 Key Observations

1. **Batch size scaling**: The noise covariance scales inversely with batch size B. Larger batches reduce noise.
2. **Position dependence**: The covariance $\Sigma_{\mathrm{emp}}(\theta, \mathcal{D}_n)$ depends on the current parameters θ, making the noise *multiplicative* rather than additive.
3. **Finite sample correction**: The factor $\frac{n-1}{n}$ in the expected covariance reflects the finite dataset size. As $n \to \infty$, we recover $\mathbb{E}[\mathrm{Cov}[\xi]] = \frac{1}{B}\Sigma_{\mathrm{pop}}(\theta)$.
4. **Sampling without replacement**: If we sample without replacement, the covariance becomes:
$$\mathrm{Cov}_{\mathcal{B}}[\xi(\theta)|\mathcal{D}_n] = \frac{n-B}{n-1} \cdot \frac{1}{B}\Sigma_{\mathrm{emp}}(\theta, \mathcal{D}_n)$$
The factor $\frac{n-B}{n-1}$ represents the finite population correction.

1.8.5 Estimating the Noise Covariance

In practice, we can estimate $\Sigma_{\text{emp}}(\theta, \mathcal{D}_n)$ by:

1. Computing multiple stochastic gradients $\{g_{\mathcal{B}_j}(\theta)\}_{j=1}^m$ at the same θ
2. Estimating: $\hat{\Sigma} = \frac{B}{m-1}\sum_{j=1}^m [g_{\mathcal{B}_j}(\theta) - \bar{g}][g_{\mathcal{B}_j}(\theta) - \bar{g}]^T$

where $\bar{g} = \frac{1}{m}\sum_{j=1}^m g_{\mathcal{B}_j}(\theta)$.

This characterization of gradient noise is fundamental for understanding the continuous-time limit of SGD, as the noise covariance $\Sigma(\theta)$ directly determines the diffusion coefficient in the limiting stochastic differential equation.

2 Continuous-time limit of SGD

2.1 Heuristic derivation of the continuous-time limit

Consider the SGG update:

$$\theta_{k+1} = \theta_k - \eta_k g_{\mathcal{B}_k}(\theta_k) \tag{12}$$

where $g_{\mathcal{B}_k}(\theta_k)$ is the gradient at the k-th iteration for batch $\mathcal{B}_k$. Define the gradient noise as:

$$\xi(\theta_k) := g_n(\theta_k) - g_{\mathcal{B}_k}(\theta_k) \tag{13}$$

where $g_n(\theta_k)$ is the empirical gradient. The new update can be written as:

$$\Delta\theta_{k+1} := \theta_{k+1} - \theta_k = -\eta_k g_{\mathcal{B}_k}(\theta_k) + \eta_k \xi(\theta_k) \tag{14}$$

The 1-sample gradient expectation is:

$$\mathbb{E}[g_i(\theta_k)] = \frac{1}{n}\sum_{i=1}^n g_i(\theta_k) = g_n(\theta_k) \tag{15}$$

Hence the expectation of the noise is zero. The 1-sample noise covariance is:

$$\Sigma(\theta_k) := \frac{1}{n}\sum_{i=1}^n [g_i(\theta_k) - g_n(\theta_k)][g_i(\theta_k) - g_n(\theta_k)]^T \tag{16}$$

where $g_i(\theta_k)$ is the gradient at the i-th data point. The batch noise covariance is:

$$\Sigma_{\text{batch}}(\theta_k) := \frac{1}{B^2}\sum_{i\in\mathcal{B}_k} Cov(\xi(\theta_k)) = \frac{1}{B}\Sigma(\theta_k) \tag{17}$$

Assume that the learning rate is independent of k. Assume that batch size is large enough to apply CLT and much smaller than the dataset size i.e. $B \ll n$ (**independent batches?**) and $\frac{B}{n} \ll 1$. Assume that number of samples is large enough to apply CLT. Assume that samples are i.i.d. and that the gradient update (and maybe gradient noise?) have finite variance. Under these assumptions, we can apply the CLT to the batch gradient and get:

$$\xi(\theta_k) \sim \mathcal{N}(0, \frac{1}{B}\Sigma(\theta_k)) \tag{18}$$

Proper scaling of the noise is crucial to apply CLT and get the continuous-time limit. Physical intuition: the average displacement of the noise is 0 but the the typical displacement of the noise is in $\sqrt{t}$ where $t = N\eta$ for N steps. More specifically

$$Var(\sum_{k=1}^{N} \eta\xi_k) = N\eta^2\Sigma(\theta_k) = t\eta\Sigma(\theta_k) \tag{19}$$

$$std(\sum_{k=1}^{N} \eta\xi_k) = \sqrt{t\eta}\sqrt{\Sigma(\theta_k)} \tag{20}$$

with $N = \frac{t}{\eta}$ for N steps. We see that as $\eta \to 0$, the typical displacement of the accumulated noise goes to 0. If we use the scaling $\sqrt{(}\eta)$ for the noise, we get:

$$Var(\sum_{k=1}^{N} \sqrt{(}\eta)\xi_k) = N\eta\Sigma(\theta_k) = t\Sigma(\theta_k) \tag{21}$$

$$std(\sum_{k=1}^{N} \sqrt{(}\eta)\xi_k) = \sqrt{t}\sqrt{\Sigma(\theta_k)} \tag{22}$$

With $N = t/\eta$ for N steps. We see that the noise as proper scaling as the typical displacement is in $\sqrt{t}$ and does not depends on η. The continuous-time limit of SGD is:

$$d\theta_t = -\nabla L_n(\theta_t)dt + \sqrt{\frac{\eta}{B}\Sigma(\theta_t)}dW_t \tag{23}$$

where W_t is a Wiener process. Note that a more rigorourous derivation would use the Lindeberg condition and the Donsker's theorem.

2.2 Key assumptions

- Gradient noise has finite variance
- η_k is independent of k
- $\eta \to dt$
- $B \ll n$ and $\frac{B}{n} \ll 1$
- No auto-correlation in the noise
- Samples are i.i.d.
- For now we are not considering momentum.
- This contrasts with Adam which has a momentum and a second order correction of past gradient.

2.3 Why do we care: the tempered posterior distribution

Using the machinery of Fokker-Planck equation, we can derive the stationary distribution of the SGD process as the tempered posterior distribution. The Fokker-Planck equation is:

$$\frac{\partial p(\theta, t)}{\partial t} = \nabla \cdot \left(\nabla L_N(\theta) p(\theta, t) + \frac{1}{2}\frac{\eta}{B}\Sigma(\theta)\nabla p(\theta, t) \right) \tag{24}$$

Assume that the noise covariance matrix is positive definite. Let the probability current be:

$$j(\theta, t) = \nabla L_N(\theta) p(\theta, t) + \frac{1}{2}\frac{\eta}{B}\Sigma(\theta)\nabla p(\theta, t) \tag{25}$$

The stationarity condition implies that the probability current is divergence free:

$$\nabla \cdot j(\theta, t) = 0 \tag{26}$$

Assume a detailed balance condition i.e. $j(\theta, t) = 0$. This implies that the stationary distribution satisfies:

$$\nabla L_N(\theta) p(\theta) + \frac{1}{2}\frac{\eta}{B}\Sigma(\theta)\nabla p(\theta) = 0 \tag{27}$$

Assume that the noise covariance matrix is constant $\Sigma(\theta) = \Sigma$, positive and definite. Then solving the latter differential equation gives us the stationary distribution as the tempered posterior distribution:

$$p_\infty(\theta) = \frac{1}{Z}\exp\left(-\beta\Sigma^{-1}L_N(\theta)\right) \tag{28}$$

where $\beta = \frac{B}{2\eta}$ is the inverse temperature. In the case where the noise covariance matrix is not constant but still positive and definite, we get:

$$p_\infty(\theta) = \frac{1}{Z}|\Sigma(\theta)|^{-1/2}\exp\left(-\frac{2B}{\eta}L_N(\theta)\right) \tag{29}$$

where Z is the partition function. In that case, the stationary distribution is given by the tempered posterior i.e. up to the temperature parameter SGD is similar to Bayesian inference. This result is been investigated in more details by Mandt. The assumptions in this derivation are not realistic since the noise covariance is singular i.e. not definite positive. In fact, deep neural networks are highly degenerate and we also know that SGD favour flat minima.

2.4 Why do we care: the tempered posterior distribution

3 TODO

- Explain why we care about continuous-time limit and connection with the tempered posterior distribution (also mention invertability of the noise covariance)
- Have a better understanding of the detailed balance condition and links with the curl of the probability current (leading to a non stationary distribution but to an orbit instead)
- Check proof of stationary distribution for the case where the noise covariance is not constant.
- Explain connection with SLT and spectrum between continuous limit with all assumptions and discrete SGD
- Include momentum considerations.

4 Appendix

4.0.1 Detailed Derivation of Batch Noise Covariance

Let's carefully work through the derivation of the batch noise covariance, addressing why we condition on the dataset.

Why Condition on the Dataset?

In SGD, there are two layers of randomness:

1. **Dataset randomness**: The dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ consists of random samples from the population

2. **Mini-batch randomness**: Given a fixed dataset, we randomly select mini-batches

When analyzing a single SGD step, the dataset is already fixed—it's the data we have. The only randomness at each iteration comes from mini-batch selection. This is why we condition on $\mathcal{D}_n$.

Setting up the Problem

Fix a dataset $\mathcal{D}_n$. For this fixed dataset, we have:

- Individual gradients: $g_i(\theta) = \nabla_\theta \ell(\theta; X_i, Y_i)$ (these are now *fixed* functions of θ)
- Empirical gradient: $g_n(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta)$ (also fixed)

The mini-batch gradient for a random batch $\mathcal{B}$ is:

$$g_{\mathcal{B}}(\theta) = \frac{1}{B}\sum_{i\in\mathcal{B}} g_i(\theta)$$

The batch noise is:

$$\nu_{\text{batch}}(\theta) = g_{\mathcal{B}}(\theta) - g_n(\theta)$$

Computing the Conditional Expectation

For sampling with replacement, each index in $\mathcal{B}$ is drawn independently and uniformly from $\{1, ..., n\}$. Let's denote these random indices as $I_1, ..., I_B$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)|\mathcal{D}_n] &= \mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(\theta) - g_n(\theta)|\mathcal{D}_n] && (30)\\
&= \mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(\theta)|\mathcal{D}_n] - g_n(\theta) && (31)\\
&= \mathbb{E}_{I_1,...,I_B}\left[\frac{1}{B}\sum_{j=1}^B g_{I_j}(\theta)\middle|\mathcal{D}_n\right] - g_n(\theta) && (32)
\end{aligned}$$

Since each I_j is uniformly distributed on $\{1, ..., n\}$:

$$\begin{aligned}
\mathbb{E}_{I_j}[g_{I_j}(\theta)|\mathcal{D}_n] &= \sum_{i=1}^n \mathbb{P}(I_j = i)\cdot g_i(\theta) && (33)\\
&= \sum_{i=1}^n \frac{1}{n}\cdot g_i(\theta) && (34)\\
&= \frac{1}{n}\sum_{i=1}^n g_i(\theta) = g_n(\theta) && (35)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)|\mathcal{D}_n] &= \frac{1}{B}\sum_{j=1}^B \mathbb{E}_{I_j}[g_{I_j}(\theta)|\mathcal{D}_n] - g_n(\theta) && (36)\\
&= \frac{1}{B}\sum_{j=1}^B g_n(\theta) - g_n(\theta) && (37)\\
&= g_n(\theta) - g_n(\theta) = 0 && (38)
\end{aligned}$$

Computing the Conditional Covariance

Now for the covariance. Since $\mathbb{E}[\nu_{\text{batch}}] = 0$:

$$\text{Cov}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)|\mathcal{D}_n] = \mathbb{E}_{\mathcal{B}}[\nu_{\text{batch}}(\theta)\nu_{\text{batch}}(\theta)^T|\mathcal{D}_n] \tag{39}$$

We have:

$$\nu_{\text{batch}}(\theta) = g_{\mathcal{B}}(\theta) - g_n(\theta) \tag{40}$$

$$= \frac{1}{B}\sum_{j=1}^{B} g_{I_j}(\theta) - g_n(\theta) \tag{41}$$

$$= \frac{1}{B}\sum_{j=1}^{B}[g_{I_j}(\theta) - g_n(\theta)] \tag{42}$$

Therefore:

$$\text{Cov}_{\mathcal{B}}[\nu_{\text{batch}}|\mathcal{D}_n] = \mathbb{E}_{\mathcal{B}}\left[\left(\frac{1}{B}\sum_{j=1}^{B}[g_{I_j}(\theta) - g_n(\theta)]\right)\left(\frac{1}{B}\sum_{k=1}^{B}[g_{I_k}(\theta) - g_n(\theta)]\right)^T \middle| \mathcal{D}_n\right] \tag{43}$$

Expanding:

$$= \frac{1}{B^2}\sum_{j=1}^{B}\sum_{k=1}^{B}\mathbb{E}_{I_j,I_k}\left[[g_{I_j}(\theta) - g_n(\theta)][g_{I_k}(\theta) - g_n(\theta)]^T|\mathcal{D}_n\right] \tag{44}$$

For sampling with replacement, I_j and I_k are independent when $j \neq k$. For $j \neq k$:

$$\mathbb{E}_{I_j,I_k}[[g_{I_j}(\theta) - g_n(\theta)][g_{I_k}(\theta) - g_n(\theta)]^T|\mathcal{D}_n] \tag{45}$$

$$= \mathbb{E}_{I_j}[g_{I_j}(\theta) - g_n(\theta)|\mathcal{D}_n] \cdot \mathbb{E}_{I_k}[g_{I_k}(\theta) - g_n(\theta)|\mathcal{D}_n]^T \tag{46}$$

$$= 0 \cdot 0^T = 0 \tag{47}$$

For $j = k$:

$$\mathbb{E}_{I_j}[[g_{I_j}(\theta) - g_n(\theta)][g_{I_j}(\theta) - g_n(\theta)]^T|\mathcal{D}_n] \tag{48}$$

$$= \sum_{i=1}^{n}\mathbb{P}(I_j = i) \cdot [g_i(\theta) - g_n(\theta)][g_i(\theta) - g_n(\theta)]^T \tag{49}$$

$$= \frac{1}{n}\sum_{i=1}^{n}[g_i(\theta) - g_n(\theta)][g_i(\theta) - g_n(\theta)]^T \tag{50}$$

$$= \Sigma_{\text{emp}}(\theta, \mathcal{D}_n) \tag{51}$$

Therefore:

$$\text{Cov}_{\mathcal{B}}[\nu_{\text{batch}}|\mathcal{D}_n] = \frac{1}{B^2}\sum_{j=1}^{B}\Sigma_{\text{emp}}(\theta, \mathcal{D}_n) \tag{52}$$

$$= \frac{1}{B^2} \cdot B \cdot \Sigma_{\text{emp}}(\theta, \mathcal{D}_n) \tag{53}$$

$$= \frac{1}{B}\Sigma_{\text{emp}}(\theta, \mathcal{D}_n) \tag{54}$$

Key Insight
The factor $\frac{1}{B}$ arises because:

- We average B independent random variables (the gradients at randomly selected indices)
- Each has the same variance Σ_{emp}
- The variance of an average of B i.i.d. random variables is $\frac{1}{B}$ times the individual variance

This is why larger batch sizes reduce gradient noise—the noise variance decreases as $O(1/B)$.

4.1 Stationary distribution when the noise covariance is not constant

5 Claude questions

Consider the SGD update:

$$\theta_{k+1} = \theta_k - \eta_k g_{\mathcal{B}_k}(\theta_k) \tag{55}$$

where $g_{\mathcal{B}_k}(\theta_k)$ is the gradient at the k-th iteration for batch $\mathcal{B}_k$. Define the gradient noise as:

$$\xi(\theta_k) := g_n(\theta_k) - g_{\mathcal{B}_k}(\theta_k) \tag{56}$$

where $g_n(\theta_k)$ is the empirical gradient. What are the conditions on n and B to apply the central limit theorem to the gradient noise?

- B should remain fixed as the learning rate $\eta \to 0$
- Typically $B \ll n$ (batch size much smaller than dataset size)

The batch size should NOT scale with $1/\eta$ $1/\eta$ The gradient noise must have finite second moments: $\mathbb{E}[\|\xi(\theta_k)\|^2|\theta_k] < \infty$ For mini-batch sampling with replacement, the covariance is: $\text{Cov}[\xi(\theta_k)|\theta_k] = \frac{1}{B} \cdot \frac{n-B}{n-1} \cdot \text{Cov}_{i\sim\text{Uniform}(1,n)}[\nabla L(\theta_k; X_i)]$ This requires individual gradients to have finite variance. If sampling with replacement: No additional constraints beyond finite variance If sampling without replacement: Need $B/n \ll 1$ $B/n \ll 1$ so that dependencies between batches are negligible