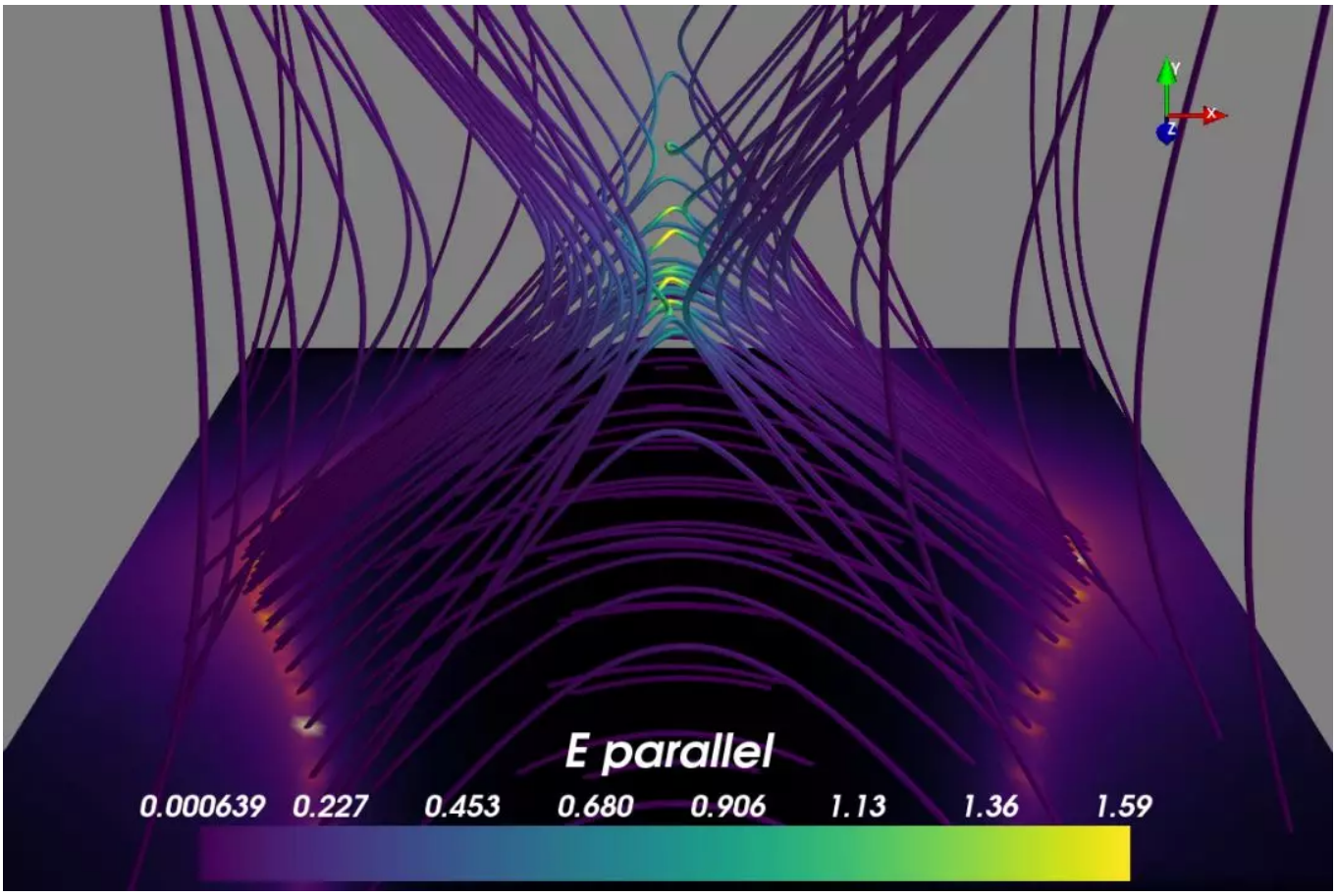
你真的了解随机梯度下降中的“全局最优”吗？

From:新智元 新智元 3/5



新智元报道

编辑：闻菲、金磊

【新智元导读】在密苏里科技大学与百度大数据实验室合作的一篇论文中，研究人员从理论视角对SGD在深度神经网络训练过程中的行为进行了刻画，揭示了SGD的随机项在其选择最终的全局极小值点的关键性作用。这项工作加深了对SGD优化过程的理解，也有助于构建深度神经网络的训练理论。

梯度下降是机器学习算法中最常用的一种优化方法。

其中，随机梯度下降 (Stochastic Gradient Descent, SGD) 由于学习速率快并且可以在线更新，常被用于训练各种机器学习和深度学习模型，很多当前性能最优 (SOTA) 模型都使用了SGD。

然而，由于SGD 每次随机从训练集中选择少量样本进行学习，每次更新都可能不会按照正确的方向进行，因此会出现优化波动。

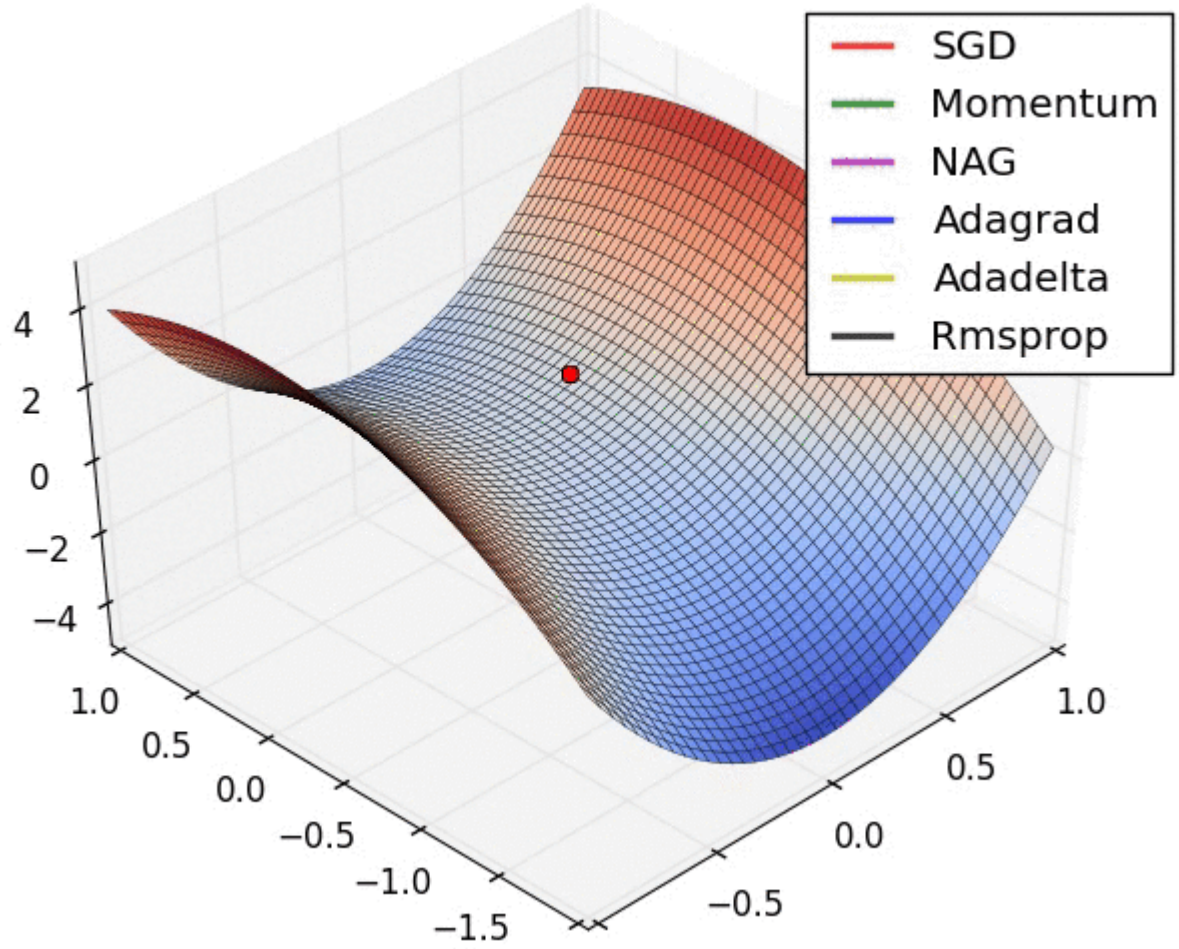
对于非凸函数而言，SGD就只会收敛到局部最优点。但同时，SGD所包含的这种随机波动也可能使优化的方向从当前的局部最优跳到另一个更好的局部最优点，甚至是全局最优。

在密苏里科技大学与百度大数据实验室日前合作公开的一篇论文中，研究人员**利用概率论中的大偏差理论对SGD在深度神经网络训练过程中的行为进行了刻画**。

“这项工作的出发点在于试图理解SGD的优化过程和GD有什么不同，尤其是SGD的随机项（也是GD所没有的）在隐式正则化中到底起到什么作用。”论文第一作者、密苏里科技大学数学系助理教授胡文清博士在接受新智元采访时说。

“通过变分分析和构造势函数，我们发现，由于有方差 (variance) 的存在，对于任何局部最优而言，SGD都有一定逃逸的可能性。”研究负责人、百度大数据实验室科学家浣军博士告诉新智元：“如果时间足够长，SGD会以马氏链的方式遍历所有的局部最优，最终达到一个全局最优。”

“对于过参数化网络 (over parameterized network)，全局最优的点在任何数据点的梯度都是0。SGD就会被限制在这样的位置上。”



不同梯度下降优化方法在损失曲面鞍点处的表现，过参数化网络的全局最优点在任何数据点的梯度都是0，SGD就会被限制在这样的位置上。

这项工作有助于我们更深刻地理解SGD在训练深度神经网络过程，以及训练其它机器学习模型中的机制和作用。

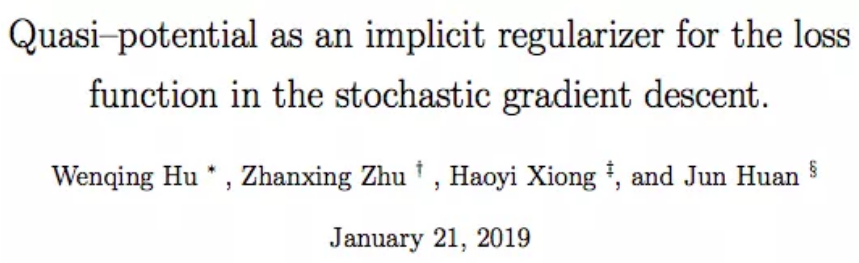
拟势函数：随机梯度下降中损失函数的隐式正则项

人们普遍认为SGD是一种“隐式正则项”，能够自己在模型或数据集中寻找一个局部最小点。

此前有研究从变分推断的角度分析SGD逃离bad minima的现象。还有研究发现，SGD的逃逸速率跟噪声协方差有关，尤其是在深度神经网络模型中。

在这篇题为《将拟势函数视为随机梯度下降损失函数中的隐式正则项》的论文中，作者提出了一种统一的方法，**将拟势作为一种量化关系的桥梁，在SGD隐式正则化与SGD的随机项的协方差结构之间建立了联系**。

“从‘拟势’这种统一的观点出发，能更清楚地从数学上描述SGD的长时间动力学。”胡文清博士说。


Quasi–potential as an implicit regularizer for the loss function in the stochastic gradient descent.

Wenqing Hu *, Zhanxing Zhu †, Haoyi Xiong ‡, and Jun Huan §

January 21, 2019

具体说，他们将随机梯度下降 (SGD) 的变分推断看做是一个势函数最小化的过程，他们将这个势函数称之为“拟势函数”(quasi–potential)，用(全局)拟势 Φ^{QP} 表示。

这个拟势函数能够表征具有小学习率的SGD的长期行为。研究人员证明，SGD最终达到的全局极小值点，既依赖于原来的损失函数f，也依赖于SGD所自带的随机项的协方差结构。

不仅如此，这项工作的理论预测对于一般的非凸优化问题都成立，揭示了SGD随机性的协方差结构在其选择最终的全局极小值点这个动力学过程的关键性作用，进一步揭示了机器学习中SGD的隐式正则化的机制。

下面是新智元对论文凸损失函数相关部分的编译，**点击“阅读原文”查看论文了解更多**。

局部拟势：凸损失函数的情况

我们假设原来的损失函数f(x)是凸函数，只允许一个最小点O，这也是它的全局最小点。设O是原点。

我们将在这一节中介绍局部准势函数，并通过哈密顿-雅可比型偏微分方程将其与SGD噪声协方差结构联系起来。分析的基础是将LDT解释为轨迹空间中的路径积分理论。

SGD作为梯度下降(GD)的一个小随机扰动

首先，我们给出一个假设：

假设1：假设损失函数f(x)允许梯度∇f(x)，即L–Lipschitz：

$$|\nabla f(x) - \nabla f(y)| \leq L|x-y| \ for\ all\ x, y \in \mathbb{R}^d \ and\ some\ L > 0 \tag{1}$$

我们假设Σ(x)是x中的分段Lipschitz，并且SDG协方差矩阵D(x)对于所有x∈$\mathbb{R}^d$是可逆的，使得：

$$TrD(x) \leq M \ for\ all\ x \in \mathbb{R}^d \ and\ some\ M > 0 \tag{2}$$

对于ε > 0，SGD过程具有接近由如下确定性方程表征的梯度下降（GD）流的轨迹：

$$\frac{dx^{\mathrm{GD}}(t)}{dt} = -\nabla f(x^{\mathrm{GD}}(t)) \ , \ x(0) = x_0 \tag{3}$$

事实上，我们可以很容易地证明有以下内容：

引理1：基于假设1，我们有，对于任何T > 0，

$$\max_{0 \leq t \leq T} \mathbf{E}|x(t) - x^{GD}(t)|^2 \leq C\varepsilon \tag{4}$$

对一些常数C = C(T, L, M) > 0。

当上述公式成立时，我们可以很容易得出在区间0≤t≤T内，x(t)和$x^{GD}(t)$收敛于$\mathcal{O}(\sqrt{\varepsilon})$。因此，在有限的时间内，SGD过程x(t)将被吸引到原点O的邻域。

由于O是凸损失函数f(x)的唯一最小点，R中的每一点都被梯度流$\mathbb{R}^d$吸引到O。

在仅有一个最小点O的情况下，也可以执行由于小的随机扰动而对吸引子(attractor)的逃逸特性的理解。

大偏差理论解释为轨迹空间中的路径积分

为了定量地描述这种逃逸特性，我们建议使用概率论中的大偏差理论(LDT)。粗略地说，这个理论给出了路径空间中的概率权重，而权重的指数部分由一个作用量泛函S给出。

局部拟势函数作为变分问题和哈密顿-雅可比方程的解

我们可以定义一个局部拟势函数为：

$$\phi_{\mathrm{loc}}^{\mathrm{QP}}(x;x_0)=\inf_{T>0}\inf_{\psi(0)=x_0,\psi(T)=x}S_{0T}(\psi)$$

(5)

将公式(5)和下面的公式6)进行结合

$$\begin{aligned}-\lim_{\varepsilon\to 0}\varepsilon\ln\rho^{\mathrm{SS}}(x) &= -\lim_{T\to\infty}\lim_{\varepsilon\to 0}\varepsilon\ln\rho(x,T|x_0,0)\\ &=\lim_{T\to\infty}\left(\inf_{\psi(0)=x_0,\psi(T)=x}S_{0T}(\psi)\right)\\ &=\inf_{T>0}\inf_{\psi(0)=x_0,\psi(T)=x}S_{0T}(\psi)\ .\end{aligned}$$

(6)

给出了平稳测度的指数渐近：

$$\rho^{\mathrm{SS}}(x)\asymp\exp\left(-\frac{1}{\varepsilon}\phi_{\mathrm{loc}}^{\mathrm{QP}}(x;x_0)\right)$$

(7)

这意味着在梯度系统只有一个稳定吸引子O的情况下，拟势 $\Phi^{QP}(x)$ 是由局部 $\Phi^{QP}_{loc}(x;\ x_0)$ 给定，这是变分问题(公式5)的解。

局部最小点的逃逸属性(根据局部拟势)

局部拟势 $\Phi^{QP}_{loc}(x;\ x_0)$ 的另一个显著特征是它描述了局部最小点的逃逸性质。从sharp极小值到flat极小值的逃逸是导致良好泛化的一个关键特征。

LDT估计提供了一种工具，可以获得退出概率的指数估计值，并从吸引子获得平均首次退出时间。

并且我们可以证明一个过程x(t)在局部最小点处的逃逸性质，如出口概率、平均逃逸时间甚至第一个出口位置，都与拟势有关。

全局拟势：SGD在各个局部极小值点之间的马氏链动力学

现在再假设损失函数f(x)是非凸的，存在多个局部极小值点。这种情况下，对每个局部极小值点的吸引区域，都可数学上构造由前述所介绍的局部拟势。

SGD在进入一个局部极小值点之后，会在其协方差结构所带来的噪声的作用下，逃逸这个局部极小值点，从而进入另一个局部极小值点。

按照前述的介绍，这种逃逸可以由局部拟势给出。然而在全局情形，不同的极小值点之间的局部拟势不一样，而从一个极小值点到另一个极小值点之间的这种由逃逸产生的跃迁，会诱导一个局部极小值点之间的马氏链。

我们的文章指出，SGD的长时间极限行为，正是以这种马氏链的方式，遍历可能的局部极小值点，最终达到一个全局极小值点。

值得一提的是，这个全局极小值点不一定是原来损失函数的全局极小值点，而是和SGD的随机性的协方差结构有关，这一点可以由上节中局部拟势的构造方式看出。

这就表明SGD的随机性所产生的协方差结构，影响了其长期行为以及最终的全局极小值点的选择。

文章中给出了一个例子，说明当损失函数f(x)有两个完全对称的全局极小值点，而其所对应的协方差结构不同的情况下，SGD会倾向于选择其中一个全局极小值点，这一个极小值点对应的协方差结构更接近各向同性(isotropic)。

未来工作

研究人员希望通过这项工作，进一步理解SGD所训练出的局部极小点的泛化性能，特别是泛化能力与协方差结构的关系。基于此，他们期待进一步的结果将不仅仅局限于overparametrized神经网络，而对一般的深度学习模型都适用。

论文链接(点击阅读原文查看)：

https://arxiv.org/pdf/1901.06054.pdf

更多阅读

- 世界首次四足后空翻，MIT机器猎豹绝杀波士顿动力！(视频)
- 30亿美金！地平线成全球估值最高AI芯片独角兽
- 1万7！华为震撼发布全球最快5G折叠屏手机，余承东炫业界最大创新