1 Optimization

Note takers: Achddou R., Al Marjani, Brogat-Motte, Foucault, Graziani, Le Corre, Pierrot, Sentenac, Yang

Instructor: Francis Bach (Inria)

1.1 Convex Optimization

1.1.1 Setup and notation

Given a dataset: $(x_i, y_i)_{1 \leq i \leq n}$ and a predictor function $f_\theta : \mathcal{X} \to \mathbb{R}$, our goal is to minimize

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta), \quad \text{where}$$

- The loss function ℓ is convex in its second argument (typically quadratic or logistic).
- $f_\theta : \mathcal{X} \to \mathbb{R}$ can be linear $\theta^T \varphi(x)$, where $\theta, \varphi(x) \in \mathbb{R}^d$ or non-linear (neural network).
- The regularizer $\Omega(\theta)$ is typically the squared L_2 or L_1 norm.

GOAL:

1. Minimize the training error $F(\theta)$, where $(x_i, y_i)_{1 \leq i \leq n}$ are i.i.d from an unknown distribution $\mathbb{P}$.
2. Control the testing error $\mathbb{E}_{(X,Y)\sim\mathbb{P}} \left[l(Y, f_{\widehat{\theta}}(X)) \right]$.
3. Do this efficiently, i.e. in $o(n)$ time.

Throughout this lecture, we restrict our attention to the case where

$$\ell(y, f_\theta(x)) = \frac{1}{2} |y - \theta^\top \varphi(x)|^2.$$

This is not a limitation, as the same techniques can be used to analyze the general case of convex optimization under smoothness assumptions.

1.1.2 Gradient Descent

We can rewrite $F(\theta) = \frac{1}{2n}\|Y - \Phi\theta\|_2^2$, where $Y \in \mathbb{R}^n$, $\theta \in \mathbb{R}^d$ and $\Phi \in \mathbb{R}^{n\times d}$. The gradient of F is given by

$$F'(\theta) = \frac{1}{n}\Phi^\top(\Phi\theta - Y) \in \mathbb{R}^d.$$

Let's look for a critical point (which is also a global minimum since the loss is strongly convex):

$$\begin{aligned} F'(\theta^\star) = 0 &\iff \Phi^\top\Phi\theta^\star = \Phi^\top Y, \\ &\iff \theta^\star = (\Phi^\top\Phi)^{-1}\Phi^\top Y, \end{aligned} \tag{1.1}$$

where we assumed that $\Phi^\top\Phi$ is invertible.

Gradient descent: Define $H := \frac{1}{n}\Phi^\top\Phi$. Then GD starts at $\theta_0 = 0$ and performs the update

$$\begin{aligned} \theta_t &= \theta_{t-1} - \gamma F'(\theta_{t-1}) \\ &= \theta_{t-1} - \gamma\left(\frac{\Phi^\top\Phi\theta_{t-1}}{n} - \frac{\Phi^\top Y}{n}\right) \end{aligned} \tag{1.2}$$

Combining equations (1.1) and (1.2) we get

$$\theta_t - \theta^\star = (I_d - \gamma H)^\top(\theta_{t-1} - \theta^\star),$$

Now we study the convergence speed of GD.
Criterion: Using the above, we have

$$\begin{aligned} F(\theta_t) - F(\theta^\star) &= \frac{1}{2}(\theta_t - \theta^\star)^\top H(\theta_t - \theta^\star) \\ &= \frac{1}{2}(\theta_0 - \theta^\star)^\top(I_d - \gamma H)^{2t}H(\theta_0 - \theta^\star) \end{aligned}$$

Wlog (it suffices to change the basis in the ambiant space by a rotation), we may assume that H is diagonal. We let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of H, $\mu := \min_i \lambda_i$ and $L := \max_i \lambda_i$.

Theorem 1.1

Assume that $\mu > 0$ (and bounded away from zero) and that $\gamma L \le 1$. Then we have a linear convergence (i.e exponential decay of the error) of Gradient Descent :

$$F(\theta_t) - F(\theta^\star) \le (1 - \gamma\mu)^{2t}[F(\theta_0) - F(\theta^\star)].$$

Clearly, the optimal stepsize is $\gamma^\star = \frac{1}{L}$, which gives a convergence speed of:

$$F(\theta_t) - F(\theta^\star) \le (1 - \frac{\mu}{L})^{2t}[F(\theta_0) - F(\theta^\star)].$$

We denote by $\kappa = \frac{L}{\mu}$ the condition number, which controls the convergence speed of GD.

Question: What if μ is very small ? Then one can write

$$H(1-2\gamma H)^{2t} = \mathrm{Diag}\left(\left(\lambda_i(1-\gamma\lambda_i)^{2t}\right)_{1\leq i\leq d} \right).$$

Lemma 1.1. *$\lambda(1-\gamma\lambda)^{2t} \leq \lambda\exp(-\gamma\lambda 2t) \leq \frac{1}{4t\gamma}$, as soon as $\gamma\lambda 2t \leq 1$.*

Theorem 1.2

Assume that $2\gamma L \leq 1$. Then we have for all $t \geq 1$, $F(\theta_t) - F(\theta^\star) \leq \frac{\|\theta_t - \theta^\star\|_2^2}{4t\gamma}$.

Remark.

1. *If F is convex, smooth and with bounded Hessians then a similar result holds.*
2. *For non-constant stepsize γ, there are also guarantees, provided that "we do not let γ be too small or too large".*
3. *GD is adaptive to the curvature of the function.*
4. *Problem 1: The complexity is not $o(n)$.*
5. *Problem 2: The convergence speed is not minimax optimal.*

1.1.3 Acceleration

Let's look at what happens when we add (Heavy ball) momentum

$$\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1}) + \delta(\theta_{t-1} - \theta_{t-2})$$

Denote $\eta_t = \theta_t - \theta^\star$. Then we have

$$\eta_t = \eta_{t-1} - \gamma H\eta_{t-1} + \delta(\eta_{t-1} - \eta_{t-2}),$$

aka a second order recursion. Based on our previous life as undergrad students, we solve the characteristic equation $r^2 = (1-\gamma\lambda)r + \delta(r-1)$. We need to make sure that the solutions are complex conjugates $\rho\exp(i\varphi)$ and $\rho\exp(-i\varphi)$. In other words, we need to ensure that

$$\begin{aligned}\Delta &= [(1-\gamma\lambda)+\delta]^2 - 4\delta \\ &= \delta^2 - 2(1+\gamma\lambda)\delta + \gamma^2\lambda^2 + 1 \leq 0.\end{aligned}$$

The roots of Δ are $\delta_{1,2} = 1 \pm \sqrt{\gamma\lambda}$. Therefore, to guarantee convergence of the sequence $(\eta_t)_{t\geq 1}$, we need to set $\delta \in [1-\sqrt{\gamma\lambda}, 1+\sqrt{\gamma\lambda}]$. We obtain convergence in $\frac{1}{t^2}$.

Convex function Nesterov Acceleration

$$\theta_t = \eta_{t-1} - \gamma F'(\eta_{t-1})$$
$$\eta_t = \theta_t + \delta(\theta_{t-1} - \theta_t)$$

Remark. *Checking homogeneity in the formulas is a quick way to make sure that the result is not awfully wrong.*

1.1.4 Stochastic Gradient Descent

We want to minimize

$$\mathbf{E}_{p(x,y)}[l(y, f_\theta(x))]$$

At each iteration, we compute

$$\theta_k = \theta_{k-1} - \gamma_k \frac{\partial l}{\partial \theta}(y_k, f_\theta(x_k)) \quad with \quad x_k, y_k \sim p(x, y)\,.$$

There are two settings:

- single pass with n iid pairs and $F(\theta)$ being the test error
- Multiple passes on finite data set with p, the empirical distance and where $F(\theta)$ is the training error

1.1.5 SGD: Least squares example

We will consider the same example of least squares in which $l(y, f_\theta(x)) = \frac{1}{2}\|y - \theta^\top \varphi(x)\|^2$, thus

$$\frac{\partial l}{\partial \theta} = \varphi(x)(\varphi(x)^\top \theta - y)$$

Assumption $y = \varphi^\top \theta_\star + \varepsilon$ with $\mathbf{E}[\varepsilon\varphi(x)] = 0$ and $\varepsilon^2 \leq \sigma^2$ a.s. and $\gamma H \leq I$

The iteration becomes with Least Squares:

$$\theta_n = \theta_{n-1} - \gamma_n \varphi(x_n)(\varphi^\top \theta - y)$$

By using $\eta_n = (\theta_n - \theta_\star)$ which is the same as GD:

$$\eta_n = \eta_{n-1} - \gamma_n \underbrace{\varphi(x_n)\varphi(x_n)^T}_{\mathbf{E}[\ldots]=H} \eta_{n-1} + \gamma_n \underbrace{\varepsilon_n \varphi(x_n)}_{\mathbf{E}[\ldots]=0}$$

Study:

$$\eta_n = \eta_{n-1} - \gamma H \eta_{n-1} + \gamma \varepsilon_n \varphi(x_n)$$

$$\eta_n = (1 - \gamma H)\eta_{n-1} + \gamma \underbrace{\varepsilon_n \varphi(x_n)}_{\mathbf{E}[\varepsilon_n \varphi(x_n)(\varepsilon_n \varphi(x_n))^T)] \leq \sigma^2 H}$$

Lemma 1.2.

$$\eta_n = \underbrace{(1 - \gamma H)^n \eta_0}_{\textit{determinstic}} + \underbrace{\gamma \sum_{k=1}^{n} (1 - \gamma H)^{n-k} \varepsilon_k \varphi(x_k)}_{\textit{zero mean}}$$

With this lemma, we can decompose $\mathbf{E}[\eta_n \eta_n^T]$ in two parts

$$\mathbf{E}[\eta_n \eta_n^T] = \underbrace{(1 - \gamma H)^n \eta_0 \eta_0^T (1 - \gamma H)^n}_{\text{Same as GD}} + \underbrace{\sum_{k=1}^{n} \gamma^2 (1 - \gamma H)^{n-k} \mathbf{E}[\varepsilon_k^2 \varphi(x_k)\varphi(x_k)^T](1 - \gamma H)^{n-k}}_{\text{SGD extra term}}$$

Our goal is to consider,

$$\mathbf{E}[F(\theta_n) - F(\theta_\star)] = \frac{1}{2}\mathbf{E}[\eta_n^T H \eta_n] = \frac{1}{2} tr(H \mathbf{E}[\eta_n \eta_n^T])$$

By replacing $\mathbf{E}[\eta_n \eta_n^T]$, we have **an extra term compared to SGD**:

$$tr(H \sum_{k=1}^{n-1} \gamma^2 \sigma^2 H (1 - \gamma H)^{2k}) \leq \sigma^2 \gamma^2 tr(H^2 (I - (I - \gamma H)^2)^{-1})$$

$$\leq \sigma^2 \gamma^2 tr(H^2 (\gamma H)^{-1})$$

$$\leq \sigma^2 \gamma tr(H)$$

Summary:

$$\mathbf{E}[F(\theta_n) - F(\theta_\star)] \leq \frac{1}{n\gamma_n} \|\eta_0\|_2^2 + \gamma_n \sigma^2 tr(H)$$

Question: How to get a convergent algo?

- γ decreasing: Brute force way or lazy way with a constant step size depending on the horizon.
- Averaging

Sweet spot: $\gamma_n = \frac{1}{\sqrt{n}}$ is a nice idea, to get $\mathbf{E}[F(\theta_n) - F(\theta_\star)] \leq \frac{1}{\sqrt{n}}(\ldots)$ But it is not homogeneous.

1.2 SGD with averaging (beginning of lecture 2)

Recall: $l_i(\theta) = (y_i - \varphi(x_i)^T\theta)^2$

Algorithm. The algorithm is the same as the classical SGD, but the final estimate is the average of the previous θ_i.

- initialize θ_0
- $\theta_n = \theta_{n-1} - \gamma \frac{\partial l_n}{\partial \theta}(\theta_{n-1})$.
- $\bar{\theta}_n = \frac{1}{n+1}\sum_{i=0}^{n}\theta_i$

Assumption.

$$y = \theta_*^T\varphi(x) + \varepsilon, \|\varphi(x)\|_2 \leq R, |\varepsilon| \leq \sigma$$

Let us prove the convergence of this algorithm.

1.2.1 Convergence proof for Least-squares

We note $\bar{\eta}_n = \bar{\theta}_n - \theta^* = \frac{1}{n+1}\sum_{k=0}^{n}\eta_k$ with $\eta_n = \theta_n - \theta^*$.

As previously:

$$\eta_n = (1-\gamma H)^n\eta_0 + \gamma\sum_{k=1}^{n}(1-\gamma H)^{n-k}\varepsilon_k\varphi(x_k)$$

so

$$\bar{\eta}_n = \underbrace{\frac{1}{n+1}\sum_{k=0}^{n}(1-\gamma H)^k\eta_0}_{\text{A: deterministic part}} + \underbrace{\frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k}(1-\gamma H)^{n-k}\varepsilon_k\varphi(x_k)}_{\text{B: noise part}}.$$

The ε_k being independent with $\mathbb{E}[\varepsilon] = 0$, we have

$$\mathbb{E}[F(\bar{\theta}_n) - F(\theta_*)] = \frac{1}{2}\mathbb{E}[\bar{\eta}_n^T H\bar{\eta}_n] = \frac{1}{2}\left(\mathbb{E}[A^T HA] + \mathbb{E}[B^T HB]\right)$$

Here, we study the deterministic and noisy term separately.

Deterministic part.

We have

$$A = \frac{1}{n+1}\sum_{k=0}^{n}(I-\gamma H)^k\eta_0 = \frac{1}{n+1}\frac{I-(I-\gamma H)^{n+1}}{I-(I-\gamma H)}\eta_0$$
$$= \frac{(\gamma H)^{-1}}{n+1}(I-(I-\gamma H)^{n+1})\eta_0. \tag{1.3}$$

So using $0 \preceq (I-\gamma H)^{n+1} \preceq I$ from $\gamma H \preceq I$, we get

$$\mathbb{E}[A^T H A] \leq \frac{\eta_0^T H^{-1}\eta_0}{\gamma^2 n^2}. \tag{1.4}$$

Noise part.

We have

$$B = \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k}(1-\gamma H)^{k-j}\varepsilon_j\varphi(x_j)$$
$$= \frac{\gamma}{n+1}\sum_{j=1}^{n}\Big(\sum_{k=j}^{n}(1-\gamma H)^{k-j}\Big)\varepsilon_j\varphi(x_j)$$
$$\approx \frac{\gamma}{n+1}\sum_{j=1}^{n}(\gamma H)^{-1}\varepsilon_j\varphi(x_j) \tag{1.5}$$

So, we get

$$\mathbb{E}[B^T H B] \leq \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}[\varphi(x_j)^T\varepsilon_j^T H^{-1}\varepsilon_j\varphi(x_j)]$$
$$\leq \frac{\sigma^2}{n^2}\sum_{j=1}^{n}Tr(H^{-1}\mathbb{E}[\varphi(x_j)\varphi(x_j)^T])$$
$$\leq \frac{\sigma^2 d}{n} \tag{1.6}$$

Summing the two parts.

From equations 1.4 and 1.6 we obtain :

$$\mathbb{E}[F(\bar{\theta}_n) - F(\theta_*)] \leq \frac{\eta_0^T H^{-1}\eta_0}{\gamma^2 n^2} + \frac{\sigma^2 d}{n} \tag{1.7}$$

1.2.2 General case SGD: averaging

In this section we prove the convergence in a more global setting than least squares.

We have :

$$\|\theta_n - \theta_*\|^2 = \|\theta_{n-1} - \theta_*\|^2 - 2\gamma(\theta_{n-1} - \theta_*)^T \frac{\partial l_n}{\partial \theta} + \gamma^2 \|\frac{\partial l_n}{\partial \theta}\|^2$$

with

$$\frac{\partial l_n}{\partial \theta} = \varphi(x_n)\varphi(x_n)^T(\theta_{n-1} - \theta_*) + \varepsilon_n \varphi(x_n)$$

and

$$\|\frac{\partial l_n}{\partial \theta}\|^2 \leq 2\left(\|\varphi(x_n)\varphi(x_n)^T(\theta_{n-1} - \theta_*)\|^2 + \|\varepsilon_n \varphi(x_n)\|^2\right).$$

So

$$\begin{aligned}
\mathbb{E}(\|\theta_n - \theta_*\|^2 | \mathcal{F}_{n-1}) &\leq \|\theta_{n-1} - \theta_*\|^2 - 2\gamma(\theta_{n-1} - \theta_*)^T H(\theta_{n-1} - \theta_*) + 2\gamma^2\sigma^2R^2 && (1.8)\\
&\quad + 2\gamma^2(\theta_{n-1} - \theta_*)^T R^2 H(\theta_{n-1} - \theta_*) \\
&\leq \|\theta_{n-1} - \theta_*\|^2 + 2\gamma^2\sigma^2R^2 - 2\gamma(1 - \gamma R^2)((\theta_{n-1} - \theta_*)^T H(\theta_{n-1} - \theta_*)) && (1.9)
\end{aligned}$$

Moreover,

$$-2\gamma(1 - \gamma R^2) \leq -\gamma$$

(from $\gamma R^2 \leq 1/2$) and

$$((\theta_{n-1} - \theta_*)^T H(\theta_{n-1} - \theta_*)) = 2(F(\theta_{n-1}) - F(\theta_*))$$

so

$$\mathbb{E}(F(\theta_{n-1}) - F(\theta_*)) \leq \frac{1}{2\gamma}(\mathbb{E}(\|\theta_{n-1} - \theta_*\|^2) - \mathbb{E}(\|\theta_n - \theta_*\|^2)) + \gamma\sigma^2R^2$$

then from Jensen inequality :

$$\mathbb{E}(F(\bar{\theta}_{n-1}) - F(\theta_*)) \leq 1/N \sum_{n=1}^{N} \mathbb{E}(F(\theta_{n-1}) - F(\theta_*)) \leq \frac{\|\theta_0 - \theta_*\|^2}{2\gamma N} + \gamma\sigma^2R \qquad (1.10)$$

Shown in Bach and Moulines [2013] : $\gamma\sigma^2R$ can be replaced by $\frac{\sigma^2 d}{n}$.

1.2.3 Dual coordinate ascent

Ref Shalev-Shwartz and Zhang [2012] **Finite sum set up.**

$F(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda/2\|\theta\|_2^2$, $y \in R^n, \Phi \in R^{n\times d}$.

Closed-form solution derivation.

The problem is stated as follows :

$$\min_{u=\Phi\theta\in R^n, \theta\in R^d} \frac{1}{2n}\|y-u\|_2^2 + \lambda/2\|\theta\|_2^2,$$

and can be re-written as :

$$\min_{u,\theta}\max_{\alpha\in R^n} \frac{1}{2n}\|y-u\|_2^2 + \lambda/2\|\theta\|_2^2 + \lambda\alpha^T(u-\Phi\theta).$$

Assuming that we have $\theta^* = \Phi^T\alpha$, we have :

$$\max_{\alpha\in R^n} -\lambda/2\alpha^T\Phi\Phi^T\alpha + \min_u \frac{1}{2n}(\|y\|^2 + \|u\|^2 - 2y^Tu) + \lambda\alpha^Tu \tag{1.11}$$

We begin by minimizing the term depending in u :

$$\begin{aligned}\min_u \frac{1}{2n}(\|y\|^2 + \|u\|^2 - 2y^Tu) + \lambda\alpha^Tu &= \min_u \frac{1}{2n}\|y\|^2 + \frac{1}{2n}\|u\|^2 - u^T(y/n - \lambda\alpha) \\ &= -\frac{n}{2}\|y/n - \lambda\alpha\|_2^2 + \frac{1}{2n}\|y\|^2 \\ &= -n\lambda^2/2\|\alpha\|_2^2 + \lambda\alpha^Ty \end{aligned} \tag{1.12}$$

as $u^* = y - n\lambda\alpha$. Replacing this term in equation 1.11, we get :

$$\max_{\alpha\in R^n} -\frac{\lambda}{2}\alpha^T\Phi\Phi^T\alpha - \frac{n\lambda^2}{2}\|\alpha\|_2^2 + \lambda\alpha^Ty = \min_\alpha G(\alpha)\lambda \tag{1.13}$$

where we define G as $G(\alpha) = -\frac{1}{2}\alpha^T\Phi\Phi^T\alpha - \frac{n\lambda}{2}\|\alpha\|_2^2 + \alpha^Ty$.

From equation 1.13 we get that :

$$\alpha_* = (\Phi\Phi^T + n\lambda I)^{-1}y$$

$\theta = \Phi^T\alpha = \sum_{i=1}^n \varphi(x_i)\alpha_i$

$\nabla^2 G(\alpha) = \Phi\Phi^T$

$\text{Diag}\nabla^2 G(\alpha) \leq R^2$

Coordinate ascent algorithm.

- Choose coordinate at random in $1, \dots, n$
- Optimize with respect to α_i

Lemma 1.3:

if $h(\beta)$ quadractic then $\inf_\beta h = h(\beta_0) - \frac{1}{2}\frac{h'(\beta_0)^2}{h''(\beta_0)}$.

What is the convergence rate?

$\mathbb{E}(G(\alpha^t) - G(\alpha_*)) = \frac{1}{n}\sum_{i=1}^n G(\alpha) - \frac{1}{2R^2}\nabla G(\alpha)_i^2 - G(\alpha_*) = G(\alpha) - \frac{1}{2nR^2}\|\nabla G(\alpha)\|^2 - G(\alpha_*)$

Losajevich condition.

If G is λ-strongly convex $G(\alpha) - G(\alpha_*) \leq \frac{1}{2\lambda}\|\nabla G(\alpha)\|$, then

$\mathbb{E}(G(\alpha^t) - G(\alpha_*)) \leq (G(\alpha) - G(\alpha_*))(1 - \frac{\lambda}{R^2+n\lambda})$

to reach ε precision

$t \approx \frac{R^2+n\lambda}{\lambda}\log 1/\varepsilon$.

1.3 Global Optimization

1.3.1 Gradient Descent for a single hidden layer: Intro

The predictor is

$$h(x) := \frac{1}{m}\theta_2^T\sigma(\theta_1^T x) = \frac{1}{m}\sum_{j=1}^m \theta_2(j)\sigma\left(\theta_1(.,j)^T x\right).$$

It is rewritten

$$h(x) = \frac{1}{m}\sum_{j=1}^m \Psi(w_j) \text{ with } \Psi(w_j)(x) := \theta_2(j)\sigma\left(\theta_1(.,j)^T x\right).$$

Goal: Minimize

$$R(h) := \mathbb{E}_{p(x,y)}\ell(y, h(x)) \text{ with } R \text{ convex.}$$

Main insight:

$$h = \int_{\mathcal{W}} \Psi(w)d\mu(w) \text{ with } d\mu(w) = \frac{1}{m}\sum_{j=1}^m \delta_{w_j}.$$

Overparametrized regime $\implies$ mean fields limit, measure μ with densities.

We want to minimize with respect to measure μ:

$$R\left(\int_{\mathcal{W}} \Psi(w)d\mu(w)\right).$$

μ is represented by a finite set of m particules, gradient descent on $(w_1, \dots, w_m)$.

Three main questions:

- Algorithm limit when m gets large
- Global conv. to a global minimum
- Prediction performance.

1.3.2 Derivations

Goal Find

$$\min_{\mu} F(\mu) = R\left(\int_{\mathcal{W}} \Psi(w) d\mu(w)\right).$$

We study Gradient flows instead of GD:

$$\dot{W} = -m\nabla F(W) \text{ with } W = (w_1, \dots, w_m).$$

It's an idealisation of this GD with small steps:

$$W_{k+1} = W_k - \gamma \nabla F(W_k).$$

Reindexing:

$$\overline{W}_{k\gamma+\gamma} = \overline{W}_{\gamma k} - \gamma \nabla F(\overline{W}_{\gamma k}).$$

Assimilating $k\gamma \leftarrow t, \gamma \leftarrow dt$ gives:

$$\overline{W}_{t+dt} = \overline{W}_t - \gamma \nabla F(\overline{W}_t).$$

This is an Euler scheme of the gradient flow differential equation.

What's the bound on the deviation between the gradient flow and the GD ? (This problem is not treated, we do computations at the limit in the course).

Note: SGD is not a Langevin diffusion in the limit as the noise is multiplied by γ. It would be if the noise term were multiplied by $\gamma^{1/2}$.

First derivation on linear Networks (no activation function).

Loss function $R : \mathcal{R}^{d\times d} \to \mathcal{R}$.

The particles' parameters $w_1, \dots, w_m$ lie in $\mathcal{R}^d$, $W = (w_1, \dots, w_m) \in \mathcal{R}^{d\times m}$.

The function to be minimized

$$F(W) = R\left(\frac{1}{m}\sum_{j=1}^{m} w_j w_j^T\right) = R\left(\frac{1}{m} W W^T\right)$$

is the composition of a convex function with a quadratic function of the parameters.

Proposition 1. *The function $M := \frac{1}{m}WW^T$ is positive semi-definite.*

The goal is to minimize a convex function over the set of SDP matrices.

Lemma 1.3. *M is optimal iff:*

$$\begin{cases} M \succcurlyeq 0 \\ \nabla R(M) \succcurlyeq 0 \\ tr(M\nabla R(M)) = 0. \end{cases}$$

Proof. **1. The condition is necessary**. The first condition is true by construction. Also, if M is optimal then:

$$\begin{aligned} \forall M + \varepsilon\Delta \succcurlyeq 0, R(M + \varepsilon\Delta) \geq R(M), \\ \implies R(M) + \varepsilon \text{tr}\Delta\nabla R(M) + o(\varepsilon) \geq R(M) \\ \implies \text{tr}\Delta\nabla R(M) \geq 0. \end{aligned}$$

For any $u \in \mathbb{R}^d$, pick $\Delta = uu^T$, this implies $u^T\nabla R(M)u \geq 0$, which implies $\nabla R(M) \succcurlyeq 0$.

Pick $\Delta = \pm M$, $\implies$ $tr(M)\nabla R(M) = 0$. Note that the convexity of R is not used to proof the necessary condition.

2. The condition is sufficient. For any matrix $N \succcurlyeq 0$:

$$\begin{aligned} R(N) &\geq R(M) + \text{tr}\nabla R(M)(N - M), \text{ (by convexity of R)} \\ &\geq R(M) + \underbrace{\text{tr}\nabla R(M)N}_{\geq 0} - \underbrace{\text{tr}\nabla R(M)M}_{=0} \\ &\geq R(M). \end{aligned}$$

□

Formulation of the problem We study the gradient flow of

$$F(W) = R\left(\frac{1}{m}WW^T\right).$$

It respects equation:

$$\dot{W} = -\frac{m}{2}\nabla F(w).$$

The speed of the gradient flow is tuned for elegance of computations. Let's compute the gradient explicitly.

$$\begin{aligned} F(W + \Delta) &= R\left(\frac{1}{m}W\Delta^T + \frac{1}{m}\Delta W^T + O(\|\Delta\|^2) + \frac{1}{m}WW^T\right) \\ &= R(\frac{1}{m}WW^T) + \text{tr}(\nabla R(M)(\frac{1}{m}(W\Delta^T + \Delta W^T)))) + +O(\|\Delta\|^2) \\ &= R(\frac{1}{m}WW^T) + \text{tr}(\Delta^T\frac{2}{m}\nabla R(M)W) + O(\|\Delta\|^2). \end{aligned}$$

Identifying the gradient we get:

$$\dot{W} = -\frac{m}{2}\nabla F(w) = -\nabla R\left(\frac{1}{m}WW^T\right)W.$$

Goal: Running gradient flow ends on a point satisfying the conditions of optimality.

Fact 1: If W has rank d at time 0 then $W(t)$ remains full rank.

Proof. We study $r := \log\det(M)$.

Let's write the ODE for $M = \frac{1}{m}WW^T$.

$$\begin{aligned}\dot{M} &= \frac{1}{m}\dot{W}W^T + \frac{1}{m}W\dot{W}^T \\ &= \frac{1}{m}\left[-\nabla R(M)WW^T - WW^T\nabla R(M)\right] \\ &= -\nabla R(M)M - M\nabla R(M).\end{aligned}$$

Note that it depends only on M. Now for the ODE of r:

$$\begin{aligned}\dot{r} &= \mathrm{tr}(M^{-1}\dot{M}) \\ &= -\mathrm{tr}(M^{-1}\nabla R(M)M + M\nabla R(M)) \\ &= -2\mathrm{tr}\left(\nabla R(M)\right)\end{aligned}$$

This implies that if $r(0)$ is defined, then r is always defined, thus M remains full rank if it is at $t = 0$. □

Assume that $M(t)$ converges:

$$M(t) \to M_\infty.$$

For general 2 layer networks, this has not been shown, in the simpler case we are considering (Linear Networks), it has been done, we skip it for this course.

Proposition 2. *M_∞ is optimal.*

Proof. The stationarity condition gives:

$$\begin{aligned}&\dot{M} = -\nabla R(M)M - M\nabla R(M) \\ \implies &-\nabla R(M_\infty)M_\infty - M_\infty\nabla R(M_\infty) = 0 \\ \implies &\mathrm{tr}(M_\infty\nabla R(M_\infty)) = 0.\end{aligned}$$

The tricky part is to show $\nabla R(M_\infty) \succcurlyeq 0$.

Assume it's not the case, i.e. $\lambda_{\min}\left(\nabla R(M_\infty)\right) < 0$.

Define set

$$K := \{z \text{ s.t. } z^T\nabla R(M_\infty)z < 0\}.$$

If there is at least one negative eigenvalue, K has a non empty interior.

We assumed that $M(t)$ converges to M_∞, there thus exists some t_0 s.t.:

$$\|M(t_0) - M_\infty\| \leq \varepsilon.$$

Consider any $y_0 \in K$. Since $W(t_0)$ has full rank, there exists some $\alpha_0 \in \mathbb{R}^m$ s.t. $y_0 = W(t_0)\alpha_0$. Define

$$z(t) := W(t)\alpha_0.$$

Note that $z(t_0) = y_0$, therefore it belongs to K by construction. This definition implies the following ODE on z:

$$\dot{z}(t) = \dot{W}(t)\alpha_0 = -\nabla R(M)W(t)\alpha_0 = -\nabla R(M)z(t).$$

This ODE implies that $z(t)$ is always well defined if $z(0)$ is. Define the shorthand $A := \nabla R(M_\infty)$.

$$\begin{aligned}
\frac{d}{dt}\left[\frac{z^TAz}{z^Tz}\right] &= \frac{\dot{z}^TAz}{z^Tz} - \frac{z^TA\dot{z}}{z^Tz} - 2\frac{z^TAz\dot{z}^Tz}{(z^Tz)^2} \\
&= -2\frac{z^T\nabla R(M)Az}{z^Tz} + 2\frac{z^TAzz^T\nabla R(M)z}{(z^Tz)^2} \\
&=_{t\to\infty} -2\frac{z^TA^2z}{z^Tz} + 2\frac{(z^TAz)^2}{(z^Tz)^2} \leq 0.
\end{aligned}$$

The last expression is non positive due to the Cauchy-Schwarz inequality. This implies that if z enters K, it never leaves. We also have:

$$\frac{d}{dt}\|z(t)\|^2 = 2\dot{z}^Tz = -2z^T\nabla R(M)z =_{t\to\infty} -2z^TAz.$$

" $\implies$ "$\|z(t)\|^2$ diverges. (The last computations are not rigorous because made in the limit, they can be made "true" involving technicalities and ε.) Divergence is impossible, as $z(t) := W(t)\alpha_0$ and $W(t)$ converges, so the hypothesis $\lambda_{\min} < 0$ is false. □

Bibliography

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). *CoRR*, abs/1306.2119, 2013. URL http://arxiv.org/abs/1306.2119.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2010.04.005. URL https://www.sciencedirect.com/science/article/pii/S1063520310000552.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Edoardo Di Napoli, Eric Polizzi, and Yousef Saad. Efficient estimation of eigenvalue counts in an interval. *Numerical Linear Algebra with Applications*, 23(4):674–692, March 2016. doi: 10.1002/nla.2048. URL https://doi.org/10.1002/nla.2048.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *arXiv preprint arXiv:1209.1873*, 2012.

David I. Shuman, Mohammad Javad Faraji, and Pierre Vandergheynst. A multiscale pyramid transform for graph signals. *IEEE Trans. Signal Process.*, 64(8):2119–2134, 2016. doi: 10.1109/TSP.2015.2512529. URL https://doi.org/10.1109/TSP.2015.2512529.