Contents

1 Smooth Convex Optimization

We are interested in unconstrained minimization of convex and smooth $f : \mathbb{R}^n \to \mathbb{R}$ given first order oracle

$$\text{minimize}_{x \in \mathbb{R}^n} \; f(x)$$

We may impose additional assumption on f, i.e. L-lipschitz, μ-strongly convex

1.1 Gradient Descent

Gradient descent achieves sublinear convergence $\mathcal{O}(\frac{1}{\epsilon})$ for $f \in \mathscr{F}_L^1$ and $\mathcal{O}(\log \frac{1}{\epsilon})$ for $f \in \mathscr{S}_{L,\mu}^1$.

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

for some stepsize $\alpha_k \geq 0$. Note $\alpha_k = \frac{1}{L}$ is the optimal stepsize.

1.1.1 Gradient Descent with Barzilai & Borwein Stepsize

Barzilai & Borwein stepsize relaxes the constraint on monotonic descent [1]. The idea is to choose α_k such that $\alpha_k g^k$ approximates the Newton update.

$$\alpha_k = \frac{\langle u^k, v^k \rangle}{\|v^k\|^2} \qquad \text{or} \qquad \alpha_k = \frac{\|u^k\|^2}{\langle u^k, v^k \rangle}$$

where

$$u^k = x^k - x^{k-1} \qquad v^k = \nabla f(x^k) - \nabla f(x^{k-1})$$

This algorithm enjoys fast empirical convergence.

1.2 Nesterov's Accelerated Gradient

Nesterov's accelerated gradient achieves lower bound for minimization of function $f \in \mathscr{S}_{L,\mu}^1$ and improves the rate for gradient descent from $\mathcal{O}(\kappa \log \frac{1}{\epsilon})$ to $\mathcal{O}(\sqrt{\kappa} \log \frac{1}{\epsilon})$. Similarly, acceleration improves convergence rate for function $f \in \mathscr{F}_L^1$ from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

1.2.1 Intuition

The following comes from Nesterov's book [2] and lecture note.

Definition. *A pair of sequences* $(\{\phi_k(x)\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty})$ *where* $\lambda_k \geq 0$ *are called the estimating sequences of the function* $f(\cdot)$ *if*

1. $\lambda_k \to 0$ *and*
2. *(**lower bound**) for any* $x \in \mathbb{R}^n$ *and for all* $k \geq 0$, $\phi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \phi_0(x)$

In addition, If we can find some sequence of points $\{x^k\}_{k=0}^{\infty}$ such that

3. ***(upper bound)*** *for any* $x \in \mathbb{R}^n$, $f(x^k) \leq \phi_k(x)$

then the rate of convergence can be derived from convergence rate of λ_k, i.e.

$$f(x^k) - f^* \leq \lambda_k \{\phi_0^* - f^*\} \to 0$$

where $\phi_k^* = \min_{x \in \mathbb{R}^n} \phi_k(x)$. Intuitively, $\phi_k(\cdot)$ are approximations for $f(\cdot)$, providing tighter and tighter bound on the optimality gap $f(x^k) - f^*$ as $\lambda_k \to 0$. In addition, from (2) and (3), we have that the sequence $\{x^k\}$ converges to the minimizer of f.

$$f(x^k) \leq \phi_k(x^*) \leq f(x^*)$$

In [2], Nesterov showed that for $f \in \mathscr{S}^1_{\mu,L}$, we can construct estimating sequences for f recursively

$$\begin{aligned} \lambda_{k+1} &= (1 - \alpha_k)\lambda_k \\ \phi_{k+1}(x) &= (1 - \alpha_k)\phi_k(x) + \alpha_k L_k(x) \\ &\text{where} \quad L_k(x) = f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + \frac{\mu}{2} \left\| x - y^k \right\|^2 \end{aligned}$$

where $\{y^k\}_{k=0}^{\infty}$ is an arbitrary sequence of points, coefficients $\{\alpha_k\}_{k=0}^{\infty}$ satisfy $\alpha_k \in (0, 1)$ and $\sum_k \alpha_k = \infty$ with $\lambda_0 = 1$ and that $\phi_0(\cdot)$ is an arbitrary convex function. Note that ϕ_k is simply a convex combination of the previous approximate ϕ_{k-1} and a quadratic lower bound L_{k-1} on f, at some carefully chosen point y^{k-1}. If we let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2} \|x - v_0\|^2$ be a quadratic function, then $\phi_k(\cdot)$ has a convenient closed form expression

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} \|x - v_k\|^2$$

where $\{\gamma_k\}, \{v_k\}, \{\phi_k^*\}$ follow certain recurrence relation detailed in [2]. Additional constraint needs to be satisfied to ensure (3) holds.

1. For (3) to hold, it must be that $f(y^k) - \frac{1}{2L} \left\| \nabla f(y^k) \right\|^2 \geq f(x^{k+1})$, which can be achieved if we obtain x^{k+1} by taking a gradient step $x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k)$ at y^k and apply descent lemma.

2. To apply the previous, we need the coefficient before $\left\| \nabla f(y^k) \right\|^2$ to agree, i.e. want α_k such that $L\alpha_k^2 = (a - \alpha_k)\gamma_k + \alpha_k \mu$.

3. Choose y^k accordingly to ensure (3) holds

By making these constraints invariant to iterative updates, we arrive at the accelerated gradient methods. In addition to the algebra tricks, there are efforts that tries to interpret what Nesterov's method is doing under the hood. For example, [3] interpreted Nesterov's accelerated method as a linear coupling of gradient descent and mirror descent . [4] showed that in the limit of small stepsizes (when taking the gradient step to obtain x^{k+1}) is equivalent to the dynamics of some continous second-order ODE.

1.2.2 The Algorithm

There are several equivalent algorithm for Nesterov's Accelerated Gradient Method. The following came from the original paper by Nesterov in 1983 [5] and later adapted to LASSO

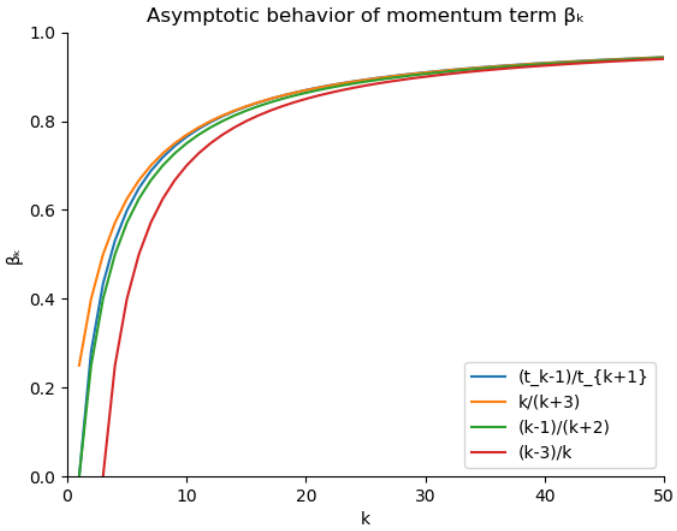
[6]. Assume $f \in \mathscr{F}_L^1$. Given $t_1 = 1$ and $y_1 = x_0$, accelerated gradient updates according to

$$
\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
t_{k+1} &= \frac{1+\sqrt{1+4t_k^2}}{2} \\
y^{k+1} &= x^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(x^{k+1} - x^k\right)
\end{aligned}
$$

We can simplify the expression by noting that (slides)

$$
\frac{t_k - 1}{t_{k+1}} = 1 - \frac{3}{k} + o(\frac{1}{k}) = \frac{k-3}{k} + o(\frac{1}{k})
$$

The momentum coefficient is asymptotically equivalent to $\frac{k-1}{k+2}$ ($\frac{t_1-1}{t_2} = 0$)



And updates is now given by

$$
\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
y^{k+1} &= x^{k+1} + \frac{k-1}{k+2}\left(x^{k+1} - x^k\right)
\end{aligned}
$$

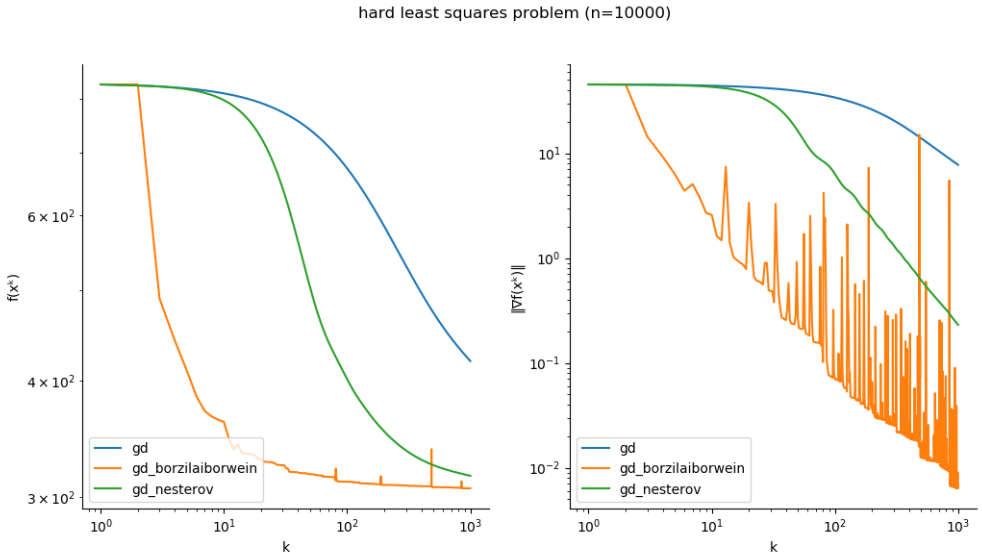
Another formulation of the algorithm comes from Nesterov's textbook [2]. If we take a constant step, i.e. $\frac{1}{L}$, to find the x^{k+1}, and that we pick $\alpha_0 = \sqrt{\frac{\mu}{L}} = 1/\sqrt{\kappa}$, which is the interpolating coefficient for recursive construction of the estimating sequence. Then we have the follwing updates

$$
\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
y^{k+1} &= x^{k+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x^{k+1} - x^k)
\end{aligned}
$$

However, in practice the condition number κ is hard to compute.

1.3 Experiments on Large Least Squares Problem

We are given a hard least squares problem of minimizing $f(x) = \frac{1}{2}\left\|D^T x - b\right\|_2^2$ where $D \in \mathbb{R}^{n\times(n+1)}$ is the differencing matrix, with all -1 on the main diagonal and all 1 on the superdiagonal. The gradient is given by $\nabla f(x) = D(D^T x - b)$. We compare gradient descent with either constant stepsize or using barzilai borwein stepsize, and nesterov's accelerated gradient descent.



We see that the barzilai borwein stepsize is the fastest method, followed by nesterov's accelerated gradient, then the naive gradient descent method.

2 Nonsmooth Convex Optimization

We are interested in constrained minimization of convex, possibly nondifferentiable, $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize}_{x\in\mathcal{C}}\ f(x)$$

given first order oracle. $\mathcal{C}$ is a simple closed convex set.

2.1 Projected Subgradient Method

Subgradient method iteratively updates as follows

$$x^{k+1} = \mathcal{P}_{\mathcal{C}}\left(x^k - \alpha_k g^k\right)$$

where $g^k \in \partial f(x^k)$ is *any* subgradient of f and that $\mathcal{P}_{\mathcal{C}}(x) = \arg\min_{y\in\mathcal{C}} \|x - y\|^2$. First order optimality condition is $\langle g(x), x - x^*\rangle \geq 0$ for any $x \in \mathcal{C}$, which is impossible to test for nontrivial function f. Therefore, using $\left\|g^k\right\| \leq \epsilon$ is not informative and subgradient method does not really have a stopping criterion.

2.1.1 Connection to Mirror Descent

Each update involves solving a subproblem of the form

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathcal{C}} \left\| x^k - \alpha_k g^k - x \right\|_2^2 \\ &= \arg\min_{x\in\mathcal{C}} \left\{ \left\| x - x^k \right\|_2^2 + 2\alpha_k \left\langle x, \nabla f(x^k) \right\rangle + \left(\alpha_k \nabla f(x^k) \right)^2 \right\} \\ &= \arg\min_{x\in\mathcal{C}} \left\{ \left\langle x, \nabla f(x^k) \right\rangle + \frac{1}{\alpha_k} D^\omega(x, x^k) \right\} \end{aligned}$$

where $D^\omega(x,y) = \frac{1}{2}\|x-y\|_2^2$ is the Bregman divergence induced by $\omega(x) = \frac{1}{2}\|x\|_2^2$. In effect, projected subgradient method is mirror descent on space endowed with ℓ-2 norm.

2.1.2 Convergence

Given bounded subgradient $\|g^k\| \le G$ and bounded domain $\|x^0 - x^*\| \le R$, subgradient method is in a sense optimal as it achieves the lower bound $\mathcal{O}(\frac{1}{\epsilon^2})$ for this problem class. The derivation as follows

$$\begin{aligned} \left\| x^{k+1} - x^* \right\|_2^2 &= \left\| \mathcal{P}_\mathcal{C}\left(x^k - \alpha_k g^k \right) - \mathcal{P}_\mathcal{C}(x^*) \right\| && \text{(Try to bound a single update)} \\ &\le \left\| x^k - \alpha_k g^k - x^* \right\|_2^2 && (\mathcal{P}_\mathcal{C} \text{ nonexpansive}) \\ &= \left\| x^k - x^* \right\|_2^2 - 2\alpha_k \left\langle g^k, x^k - x^* \right\rangle + \alpha_k^2 \left\| g^k \right\|_2^2 \\ &\le \left\| x^k - x^* \right\|_2^2 - 2\alpha_k \left(f(x^k) - f(x^*) \right) + \alpha_k^2 \left\| g^k \right\|_2^2 \\ \left\| x^{k+1} - x^* \right\|_2^2 &\le \left\| x^1 - x^* \right\|_2^2 - 2\sum_{t=1}^k \alpha_t \left(f(x^t) - f(x^*) \right) + \sum_{t=1}^k \alpha_t^2 \left\| g^t \right\|_2^2 && \text{(Telescope)} \end{aligned}$$

Then rearrange, and bound

$$2\sum_{t=1}^k \left(f(x^t) - f(x^*) \right) \le R^2 + G^2 \sum_{t=1}^k \alpha_t^2 \quad \Rightarrow \quad \min_{t\in[k]} f(x^t) - f(x^*) \le \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2\sum_{t=1}^k \alpha_t}$$

We note that $\min_{t\in[T]} f(x^t) - f(x^*) \to 0$ if stepsize is square summable but not summable, i.e. $\sum_k \alpha_k^2 < \infty$ and $\sum_k \alpha_k = \infty$. The choice of stepsize $\alpha_k = \frac{R}{\sqrt{k+1}}$ yield $\min_{t\in[k]} f(x^t) - f(x^*) = \mathcal{O}(\frac{1}{\epsilon^2})$. (3.2.3 in [2])

2.1.3 Solving Support Vector Machine w/ Subgradient Method

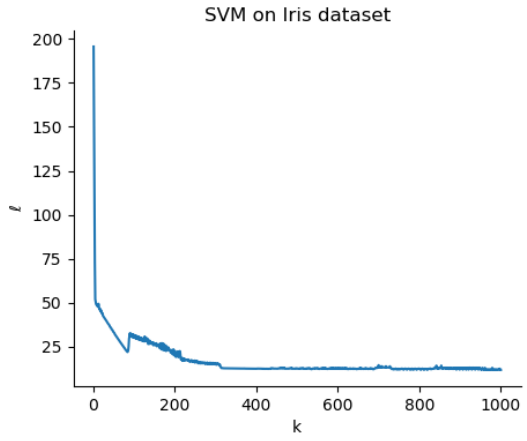
We are given data $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n \; y_i \in \{\pm 1\}\}$, support vector machine is supervised learning model that tries to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that the empirical risk and regularizer on w is minimized

$$\text{minimize}_{w,b} \quad \frac{1}{2}\|w\|_2^2 + \lambda \sum_{i=1}^m \max\left[0, 1 - y_i(w^T x_i + b)\right] \quad (:= f(w,b))$$

Support vector machines can be solved using subgradient method. We first find a subgradient of f

$$g_w^k = w^k - \lambda \sum_{i \in [m]: y_i(w^T x_i + b) < 1} y_i x_i$$
$$g_b = -\lambda \sum_{i \in [m]: y_i(w^T x_i + b) < 1} y_i$$

where we haved picked $0 \in \partial(\max 0, 1 - y_i(w^T x_i + b))$ when $y_i(w^T x_i + b) = 1$, the only case where the *max term* is non-differentiable. When tested on the Iris dataset, subgradient method worked!



2.2 Mirror Descent

3 Stochastic Optimization

We are interested in constrained minimization of $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize}_{x \in \mathcal{X}} \ [f(x) = \mathbb{E}[F(x, \xi)]]$$

where $\mathcal{X} \subset \mathbb{R}^n$ is closed, bounded convex set. ξ is a random variable, and $F(\cdot, \xi)$ is convex for all $\xi \in \Xi$, and therefore $f(\cdot)$ is convex. For uniform p_ξ over finite alphabets of size n, the problem reduces to finite sum problem

$$\text{minimize}_{x \in \mathcal{X}} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right]$$

Assume we can

1. Sample $\xi_1, \xi_2, \cdots \overset{i.i.d.}{\sim} p_\xi$
2. Given $(x, \xi) \in \mathcal{X} \times \Xi$, a first order oracle that returns a subgradient vector $G(x, \xi) \in \partial_x F(x, \xi)$. We also assume that G is unbiased, i.e. $g(x) := \mathbb{E}[G(x, \xi)] \in \partial f(x)$

3.1 Stochastic Gradient Method

We can show that if $f \in \mathscr{S}^1_{L,\mu}$, the choice of $\alpha_k = \mathcal{O}(\frac{1}{k})$ yields sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon})$ for last iterates. If $f \in \mathscr{F}^1_L$, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields a sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon^2})$ for average iterates. Stochastic gradient method (or Stochastic Approximation (SA) algorithms) has updates of the form

$$x^{k+1} = \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k G(x^k, \xi_k)\right)$$

where $\alpha_k > 0$ are stepsizes, $\mathcal{P}_{\mathcal{X}}(y) = \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|x-y\|_2^2$ is the euclidean projection onto a convex set. It is important to note that the current iterate x^k are functions of random variables $x^k := x^k(\xi_{[k-1]})$ where $\xi_{[k-1]} = (\xi_1, \cdots, \xi_{k-1})$, and therefore are random variables themselves. In addition, $x^k \perp\!\!\!\perp \xi_k$.

3.2 Convergence

Derivations copied from [7], [8] and slides. We assume

1. bounded variance for stochastic subgradient, $\mathbb{E}_\xi\left[G(x,\xi)\right] \le M^2$ given $x \in \mathcal{X}$.
2. bounded $\mathcal{X}$ where radius given by $D_{\mathcal{X}} = \max_{x\in\mathcal{X}} \|x - x^*\|_2$.

We first derive some preliminary results. Using iterated expecatation, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle G(x^k,\xi_k), x^k - x^*\right\rangle\right] &= \mathbb{E}_{\xi_{[k-1]}}\left[\mathbb{E}_{\xi_k}\left[\left\langle G(x^k(\xi_{[k-1]}),\xi_k), x^k(\xi_{[k-1]}) - x^*\right\rangle\right] \mid \xi_{[k-1]}\right] \\
&= \mathbb{E}_{\xi_{[k-1]}}\left[\left\langle \mathbb{E}_{\xi_k}\left[G(x^k(\xi_{[k-1]}),\xi_k) \mid \xi_{[k-1]}\right], x^k(\xi_{[k-1]}) - x^*\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] \qquad (1)
\end{aligned}$$

where the expectation is taken w.r.t $\xi_{[k-1]}$. We first derive a bound on residual $R_k^2 = \left\|x^k - x^*\right\|_2^2$ and expected residual $r_k^2 = \mathbb{E}\left[R_k^2\right]$ for a single update,

$$\begin{aligned}
R_{k+1}^2 &= \left\|x^k - x^*\right\|^2 \\
&= \left\|\mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k G(x^k,\xi_k)\right) - \mathcal{P}_{\mathcal{X}}(x^*)\right\|^2 && (x^* \text{ is fixed point of } \mathcal{P},\ \mathcal{P}_{\mathcal{X}}(x^*) = x^*) \\
&\le \left\|x^k - \alpha_k G(x^k,\xi_k) - x^*\right\|^2 && (\text{Nonexpansive of } \mathcal{P},\ \|\mathcal{P}_{\mathcal{X}}(x') - \mathcal{P}_{\mathcal{X}}(x)\| \le \|x' - x\|) \\
&\le R_k^2 - 2\alpha_k\left\langle G(x^k,\xi_k), x^k - x^*\right\rangle + \alpha_k^2\left\|G(x^k,\xi_k)\right\|^2 \\
r_{k+1}^2 &\le r_k^2 - 2\alpha_k\mathbb{E}\left[\left\langle G(x^k,\xi_k), x^k - x^*\right\rangle\right] + \alpha_k^2\mathbb{E}\left[\left\|G(x^k,\xi_k)\right\|^2\right] && (\text{Expectation w.r.t. } \xi_{[k]}) \\
&= r_k^2 - 2\alpha_k\mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] + \alpha_k^2 M^2 && (\text{By (1) and bounded variance})
\end{aligned}$$

3.2.1 Strongly Convex Case

If $f \in \mathscr{S}^1_{L,\mu}$, using (25), we have

$$r_{k+1}^2 \le r_k^2 - 2\alpha_k\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] + \alpha_k^2 M^2 = (1 - 2\mu\alpha_k)r_k^2 + \alpha_k^2 M^2$$

If we choose $\alpha_k = \theta/(k+1)$, where $\theta > 1/(2\mu)$. It could be shown by induction that [7]

$$r_k^2 \leq \frac{c_\theta}{k+1} \qquad \text{where} \qquad c_\theta = \max\left\{\frac{2\theta^2 M^2}{2\mu\theta - 1}, r_0\right\}$$

By (9), we derive bound on the objective value

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq \frac{1}{2}L\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \leq \frac{Lc_\theta}{2(k+1)}$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\epsilon})$ yields last iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon})$

3.2.2 Convex Case

[7] indicates that we need to increase the stepsize ($\mathcal{O}(\frac{1}{k})$ to $\mathcal{O}(\frac{1}{\sqrt{k}})$)) to ensure faster convergence rate for general convex problems, at the cost of *more noisy* trajectory. To suppress the noise, we use average iterates $\{x^k\}$ rather than last iterates as solution to the problem.

$$\begin{aligned}
r_{k+1}^2 &\leq r_k^2 - 2\alpha_k\mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] + \alpha_k^2 M^2 \\
2\alpha_k\mathbb{E}\left[f(x^k) - f(x^*)\right] &\leq 2\alpha_k\mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] \leq r_k^2 - r_{k+1}^2 + \alpha_k^2 M^2 && \text{(By 14)} \\
\sum_{i=1}^k \left(2\alpha_i\mathbb{E}\left[f(x^i) - f(x^*)\right]\right) &\leq \sum_{i=1}^k \left(r_i - r_{i+1} + \alpha_i M^2\right) = r_1^2 + M^2\sum_{i=1}^k \alpha_i^2 && \text{(Telescope)} \\
\sum_{i=1}^k \gamma_i\mathbb{E}\left[(f(x^i) - f(x^*))\right] &= \mathbb{E}\left[\sum_{i=1}^k \gamma_i(f(x^i) - f(x^*))\right] \leq \frac{r_1^2 + M^2\sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} && (/2\textstyle\sum_i \alpha_i)
\end{aligned}$$

where $\gamma_i = \alpha_i/\sum_i \alpha_i$. Let $\tilde{x}^k = \sum_{i=1}^k \gamma_i x^i$. $f(\tilde{x}^k) \leq \sum_i \gamma_i f(x^i)$ by convexity of f. Then,

$$\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] \leq \frac{r_1^2 + M^2\sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}$$

We derive tighest bound by finding minimal value of $\alpha_k = \alpha$ of the bound.

$$\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}} \qquad \alpha_k = \frac{D_{\mathcal{X}}}{M\sqrt{k}}$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields average iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon^2})$

4 Second Order Methods

For unconstrained problem $\min_{x\in\mathbb{R}^n} f(x)$, the standard Newton scheme updates according to

$$x^{k=1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k) \tag{2}$$

where $f \in C_L^{2,2}(\mathbb{R}^n)$. The method has quadratic local convergence rate when initial iterate is close to the optimum of f [2]. Cubic regularized Newton's Method converges globally to second order stationary points ($\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$) assuming $f \in C_L^{2,2}$, i.e.

twice continuously differentiable with lipschitz continuous hessian [9, 10]. The idea is to iteratively minimize a global upper bound of the objective,

$$x^{k+1} = \arg\min_{y \in \mathbb{R}^n} \tilde{f}_{x^k} y \tag{3}$$

where $\tilde{f}(y)$ is a cubic regularized quadratic model of the objective,

$$\tilde{f}_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(x)(y-x), y - x \right\rangle + \frac{L}{6} \|y - x\|^3 \tag{4}$$

This modified Newton step ensures that function values of iterates are monotonic non-increasing. Cubic regularized Newton's Method converges globally to second order stationary points ($\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$) assuming $f \in C_L^{2,2}$, i.e. twice continuously differentiable with lipschitz continuous hessian [9, 10]. The method has quadratic global convergence rate when initial iterate is close to the optimum of f [2]. Under weak non-degeneracy assumption of the Hessian matrix, the local convergence rate is super-linear of the order $\frac{4}{3}$ or $\frac{3}{2}$.

5 Minimax Optimization

5.1 Convex-Concave Minimax

Let $\mathcal{X}, \mathcal{Y}$ be nonempty set, $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. To simplify notation, let $f(\cdot) = \max_{y \in \mathcal{Y}} \phi(\cdot, y)$ and $g(\cdot) = \min_{x \in \mathcal{X}} \phi(x, \cdot)$. In general, we are intersted in the minimax problem of the form.

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y)$$

In particular, we want to find a saddle point, a pair $(x^\star, y^\star)$ where

$$x^\star \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y) = \arg\min_{x \in \mathcal{X}} f(x)$$
$$y^\star \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y) = \arg\max_{y \in \mathcal{Y}} g(y)$$

Definition. ***(Weak Minimax)*** *For any $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$, we have*

$$\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y) \leq \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y)$$

A good way to remember this is to note the follower (inner optimization) has advantage since it knows what the leader's (outer optimization) strategy. So when the follower wants to maximize the objective, it often achieves this.

Definition. ***(Strong Minimax)*** *If $\phi(\cdot, y)$ convex for all $y \in \mathcal{Y}$ and $\phi(x, \cdot)$ concave for all $x \in \mathcal{X}$, we have equality*

$$\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y) = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y)$$

5.2 Mirror Descent

5.3 Non-Convex-Non-Concave Minimax

Let $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$ be the iterates, $f^k = f(\mathbf{z}^k)$ be function evaluated at current iterates, and

$$\mathbf{H}^k = \mathbf{J}(\nabla f^k) = \begin{bmatrix} \mathbf{H_{xx}} & \mathbf{H_{xy}} \\ \mathbf{H_{yx}} & \mathbf{H_{yy}} \end{bmatrix}$$

be block-wise hessian.

Competitive Gradient Descent (CGD) Competitive Gradient Descent [11] is an generalized gradient descent algorithm for the general-sum two player game, which we will specialize to zero-sum game. Each iteration involves solving a a quadratic regularized bilinear game that approximates the general game at the current iterate.

$$\min_{\boldsymbol{\delta}_\mathbf{x}} \max_{\boldsymbol{\delta}_\mathbf{y}} \left[F^k(\boldsymbol{\delta}_\mathbf{x}, \boldsymbol{\delta}_\mathbf{y}) := \boldsymbol{\delta}_\mathbf{x}^T \nabla_\mathbf{x} f^k + \boldsymbol{\delta}_\mathbf{y}^T \nabla_\mathbf{y} f^k + \frac{1}{2} \boldsymbol{\delta}_\mathbf{x}^T \mathbf{H}_{\mathbf{x},\mathbf{y}} \boldsymbol{\delta}_\mathbf{y} + \frac{1}{2\eta} \|\boldsymbol{\delta}_\mathbf{x}\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{\delta}_\mathbf{y}\|_2^2 \right]$$

Finding 1st order local nash equilibrium involves solving a system of equations given by

$$0 = \nabla_{\boldsymbol{\delta}_\mathbf{x}} F^k = \nabla_\mathbf{x} f^k + \mathbf{H}_{\mathbf{xy}}^k \boldsymbol{\delta}_\mathbf{y} + \frac{1}{\eta} \boldsymbol{\delta}_\mathbf{x}$$
$$0 = \nabla_{\boldsymbol{\delta}_\mathbf{y}} F^k = -\nabla_\mathbf{y} f^k - \mathbf{H}_{\mathbf{yx}}^k \boldsymbol{\delta}_\mathbf{x} + \frac{1}{\eta} \boldsymbol{\delta}_\mathbf{y}$$

which emits closed form equations for $\boldsymbol{\delta}_\mathbf{x}, \boldsymbol{\delta}_\mathbf{y}$, giving rise to update of the form,

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{y}^{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^k \\ \mathbf{y}^k \end{bmatrix} - \eta \begin{bmatrix} \mathbf{I} & \eta \mathbf{H}_{\mathbf{xy}}^k \\ -\mathbf{H}_{\mathbf{yx}}^k & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \nabla_\mathbf{x} f^k \\ -\nabla_\mathbf{y} f^k \end{bmatrix}$$

Note solving for local nash of a full quadratic Taylor approximation of the game at current iterates (include terms involving $\mathbf{H_{xx}}, \mathbf{H_{yy}}$) recovers damped and regularized Newton's method. In the paper, the author uses computes approximate matrix inverse to compute the updates.

CGD with cubic regularization We can apply per-player cubic regularization,

$$\min_{\boldsymbol{\delta}_\mathbf{x}} \max_{\boldsymbol{\delta}_\mathbf{y}} \left[F^k(\boldsymbol{\delta}_\mathbf{x}, \boldsymbol{\delta}_\mathbf{y}) := \boldsymbol{\delta}_\mathbf{x}^T \nabla_\mathbf{x} f^k + \boldsymbol{\delta}_\mathbf{y}^T \nabla_\mathbf{y} f^k + \frac{1}{2} \boldsymbol{\delta}_\mathbf{x}^T \mathbf{H}_{\mathbf{x},\mathbf{y}} \boldsymbol{\delta}_\mathbf{y} + \frac{L}{6} \|\boldsymbol{\delta}_\mathbf{x}\|_2^3 - \frac{L}{6} \|\boldsymbol{\delta}_\mathbf{y}\|_2^3 \right]$$

It is not possible to write analytic equation for optimal solution. Instead, the subproblem can be computed using first order gradient methods with guaranteed on convergence to local saddle points, e.g. extra-gradient [12], concensus optimization [13].

Follow the Ridge (FR) Gradient Descent/Ascent (GDA) fails to converge with any constant learning rate. *Follow-the-Ridge* modifies gradient descent-ascent by applying an asymmetric correction term on the leader's gradient step, which encourage players to stay on the ridge of the loss surface. The approach is proved to converge and only converge to *local minimax* under mild assumptions (f twice differentiable, thrice differentiable at critical points, $\mathbf{H_{yy}}$ is invertible)

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{y}^{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^k \\ \mathbf{y}^k \end{bmatrix} - \begin{bmatrix} \eta_\mathbf{x} \mathbf{I} & \mathbf{0} \\ -\eta_\mathbf{x} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \eta_\mathbf{y} \mathbf{I} \end{bmatrix} \begin{bmatrix} \nabla_\mathbf{x} f^k \\ \nabla_\mathbf{y} f^k \end{bmatrix}$$

6 Cheat Sheet

6.1 Lipschitz Continuous

Definition 1. *$f \in C_L^{k,p}(Q)$ is k times continuously differentiable on Q if for all $x, y \in Q$,*

$$\|\nabla^p f(y) - \nabla^p f(x)\| \leq L \|y - x\| \tag{5}$$

$f \in C_L^{1,1}(\mathbb{R}^n)$ is continously differentiable on $\mathbb{R}^n$ if for all $x, y \in \mathbb{R}^n$

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\| \tag{6}$$

Definition 2. *$f \in C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$ if for all $x \in \mathbb{R}^n$, either condition is satisfied*

$$\left\|\nabla^2 f(x)\right\| \leq L \tag{7}$$

$$-L\mathbf{I} \preceq \nabla^2 f(x) \preceq L\mathbf{I} \tag{8}$$

Property for $f \in C_L^{1,1}(\mathbb{R}^n)$,

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2 \tag{9}$$

$$f\left(x - \alpha \nabla f(x)\right) \leq f(x) - \alpha \left(1 - \frac{\alpha}{2} L\right) \|\nabla f(x)\|^2 \tag{10}$$

$$f\left(x - \frac{1}{L} \nabla f(x)\right) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2 \tag{11}$$

Note (9) implies that quadratic functions $\phi_-(\cdot), \phi_+(\cdot)$ are global lower/upper bound of $f(\cdot)$ respectively, i.e. $\phi_-(y) \leq f(y) \leq \phi_+(y)$ for any $x \in \mathbb{R}^n$, where

$$\phi_-(y) = f(x) + \langle \nabla f(x), y - x \rangle - \frac{L}{2} \|y - x\|^2$$

$$\phi_+(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

Note (11) is important in proving convergence of descent methods, where we see progress made in reducing function value of iterates by steping in $-\frac{1}{L}\nabla f(x)$ is at least some constant times the gradient norm.
Property for $f \in C_M^{2,2}(\mathbb{R}^n)$,

$$\left\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\right\| \leq \frac{M}{2} \|y - x\|^2 \tag{12}$$

$$\nabla^2 f(x) - C \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + C \quad \text{where} \quad c = M \|y - x\| \mathbf{I} \tag{13}$$

6.2 Convex

Definition 3. *The following are equivalent*

1. *A continuously differentiable function f is convex on convex set Q ($f \in \mathscr{F}^1(Q)$)*
2. *For all $x, y \in Q$*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{14}$$

3. *For all* $x, y \in Q$ *and* $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{15}$$

4. *For all* $x, y \in Q$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0 \tag{16}$$

Definition 4. *A twice differentiable function* f *belongs to* $\mathscr{F}^2(Q)$ *if for any* $x \in Q$,

$$\nabla^2 f(x) \succeq 0 \tag{17}$$

6.3 Smooth & Convex

Definition 5. $f \in \mathscr{F}_L^{1,1}(Q, \|\cdot\|)$ *if* f *is convex with Lipschitz continous gradient, i.e. for all* $x, y \in Q$,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \tag{18}$$

Property of $f \in \mathscr{F}^{1,1}(\mathbb{R}^n, \|\cdot\|)$. Let $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|x - y\|^2 \tag{19}$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2 \leq f(y) \tag{20}$$

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \tag{21}$$

$$0 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \|x - y\|^2 \tag{22}$$

Note (19) implies a tighter (linear) lower bound to $f(\cdot)$ if we assume convexity. In fact, the lower bound can be improved further to a upward quadratic by (20).

6.4 Strongly Convex

Definition 6. *A continuously differentiable function* f *is strongly convex on* $\mathbb{R}^n$ *(*$f \in \mathscr{S}_\mu^1(Q, \|\cdot\|)$*) if there exists a convexity parameter* $\mu > 0$ *such that for all* $x, y \in Q$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \tag{23}$$

Property for $f \in \mathscr{S}_\mu^1(\mathbb{R}^n)$. Let $x, y \in Q$ and $\lambda \in [0,1]$,

$$f(x) \geq f(x^*) + \frac{\mu}{2} \|x - x^*\|^2 \quad \text{where} \quad \nabla f(x^*) = 0 \tag{24}$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2 \tag{25}$$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\frac{\mu}{2} \|x - y\|^2 \tag{26}$$

6.5 Smooth & Strongly Convex

Definition 7. *A continuously differentiable function* f *that is strongly convex with* L*-lipschitz continuous gradients (*$f \in \mathscr{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$*). Note* $\kappa = L/\mu \geq 1$ *is the condition number of* f.

Property for $f \in \mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$. For any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2 \tag{27}$$

References

[1] Jonathan Barzilai and Jonathan M. Borwein. "Two-Point Step Size Gradient Methods". In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1, 1988). Publisher: Oxford Academic, pp. 141–148. ISSN: 0272-4979. DOI: 10.1093/imanum/8.1.141. URL: https://academic.oup.com/imajna/article/8/1/141/802460 (visited on 03/25/2020).

[2] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. 2004. ISBN: 978-1-4020-7553-7. URL: https://dial.uclouvain.be/pr/boreal/object/boreal:116858 (visited on 03/27/2020).

[3] Zeyuan Allen-Zhu and Lorenzo Orecchia. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *arXiv:1407.1537 [cs, math, stat]* (Nov. 7, 2016). arXiv: 1407.1537. URL: http://arxiv.org/abs/1407.1537 (visited on 03/24/2020).

[4] Weijie Su, Stephen Boyd, and Emmanuel J. Candes. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *arXiv:1503.01243 [math, stat]* (Oct. 27, 2015). arXiv: 1503.01243. URL: http://arxiv.org/abs/1503.01243 (visited on 03/28/2020).

[5] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate O(1/k2)." In: *Soviet Mathematics Doklady* 27 ((2) 1983), pp. 372–376.

[6] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (Jan. 2009), pp. 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542. URL: http://epubs.siam.org/doi/10.1137/080716542 (visited on 03/27/2020).

[7] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (Jan. 1, 2009). Publisher: Society for Industrial and Applied Mathematics, pp. 1574–1609. ISSN: 1052-6234. DOI: 10.1137/070704277. URL: https://epubs.siam.org/doi/abs/10.1137/070704277 (visited on 04/01/2020).

[8] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *arXiv:1606.04838 [cs, math, stat]* (Feb. 8, 2018). arXiv: 1606.04838. URL: http://arxiv.org/abs/1606.04838 (visited on 04/01/2020).

[9] Yurii Nesterov. "Cubic Regularization of Newton's Method for Convex Problems with Constraints". In: 2006. DOI: 10.2139/ssrn.921825.

[10] Yurii Nesterov and Boris Polyak. "Cubic regularization of Newton method and its global performance". In: *Math. Program.* 108 (Aug. 1, 2006), pp. 177–205. DOI: 10.1007/s10107-006-0706-8.

[11] Florian Schäfer and Anima Anandkumar. "Competitive Gradient Descent". In: *arXiv:1905.12103 [cs, math]* (Oct. 10, 2019). arXiv: 1905.12103. URL: http://arxiv.org/abs/1905.12103 (visited on 02/28/2020).

[12] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. "A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach". In: *arXiv:1901.08511 [cs, math, stat]* (Sept. 5, 2019). arXiv: 1901.08511. URL: http://arxiv.org/abs/1901.08511 (visited on 05/10/2020).

[13] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. “The Numerics of GANs”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 1825–1835. URL: http://papers.nips.cc/paper/6779-the-numerics-of-gans.pdf (visited on 03/17/2020).